Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (36^{th}cycle)

Egocentric Video Understanding across Modalities and Domains

By

Chiara Plizzari

******

Supervisor:
Prof. Barbara Caputo

Politecnico di Torino
2024

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Chiara Plizzari
2024

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

To all the amazing people I met throughout my PhD journey.

Abstract

With the growing popularity of wearable cameras, egocentric vision has become an increasingly researched area. This perspective offers a direct view from the wearer's perspective, enabling a more direct study of human behavior. However, the introduction of these devices also presents unique challenges not encountered with traditional stationary cameras.

The goal of this thesis is to explore how multi-sensory information can address the complexities of egocentric videos. Wearable devices face significant changes in illumination, perspective, and environment, causing action recognition models to depend heavily on their training environments and struggle with generalization to new ones. In the first part of the thesis, we explore solving auxiliary tasks across various information channels from videos to enhance robustness across domains. By integrating RGB data with audio and motion information from optical flow via an auxiliary loss to align feature norms, we demonstrate that the resulting models are more generalizable and perform reliably in unseen environments. We then introduce a method using cross-instance video reconstruction through language to learn robust features against a *scenario shift*, where the same action occurs in different activities, and a *location shift*, where videos are from varied geographical locations. To this end, we curated ARGO1M, the largest dataset for action recognition generalization, containing over 1 million video clips. Our findings indicate that textual guidance significantly enhances model performance in unseen scenarios and locations.

In the second part of the thesis, we analyze previously unexplored modalities within egocentric vision. Event cameras, with their high pixel bandwidth, dynamic range, low latency, and low power consumption, effectively address challenges like fast camera motion and background clutter. We introduce N-EPIC-Kitchens, the first dataset for studying event-based data in this domain. Results demonstrate that event data perform competitively compared to traditional RGB and optical flow modalities. Finally, we integrate 3D scene information with appearance-based models to overcome the limitations of 2D images' narrow field of view and incomplete scene views. We introduce the task "Out of Sight, Not Out of Mind", which involves tracking object locations around the user over time, even when not

visible, using both frame-based images and 3D object positioning. Our findings show that 3D information significantly enhances the capability of egocentric vision systems to fully capture and understand the surrounding context.

Throughout this thesis, we highlight the importance of utilizing information from multiple channels and demonstrate that focusing on these aspects can significantly improve egocentric video understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis explores multi-modal egocentric video understanding, focusing on domain generalization in the first part and the adoption of new modalities in egocentric vision in the second part. This chapter outlines the goals, motivations, and contributions of our work.

1.1 Goals

Creating tools that support human activities, enhance quality of life, and boost our ability to achieve our desires has always been a fundamental goal for humanity. Among these innovations, digital computing has profoundly transformed our history, with mobile technology playing a pivotal role. Today, smartphones have become essential for outdoor navigation, recording life's moments, and connecting us with both old and new experiences. Yet, there is a growing interest for the next evolution in mobile tech: *wearable computing*. This concept, often depicted in movies, fiction, and pop culture, represents a significant step forward in how we envision our interaction with technology[1]. Wearable cameras enable the collection of visual information from a human perspective. Analyzing this data through egocentric (first-person) vision offers a more direct approach to study human behavior. In egocentric videos, camera movements are typically driven by the wearer's intentions and activities, with manipulated objects usually clearly visible in the frame. This direct correlation between the camera wearers' viewpoint and their interactions offers a unique, first-person perspective that enhances the understanding of human behavior and task execution. Additionally, the clear

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

visibility of manipulated objects not only aids in studying human-object interactions but also provides valuable context for interpreting the user's immediate environment and actions. This unique viewpoint has already found many applications in assistive technologies OhnBar et al. (2018), robotics (Park et al., 2016), entertainment (Liang et al., 2015; Taylor et al., 2020) and autonomous vehicles (Hirakawa et al., 2018).

However, the transition from traditional, stationary third-person cameras to the dynamic, first-person perspective offered by wearable cameras introduces a range of challenges. The rapid movement of the camera often results in videos affected by motion blur, and the user's hands or arms frequently occlude the camera's view. Additionally, the camera's narrow field of view restricts it to capture only partial observations of the scene. Furthermore, as the camera is worn by the user, wearable devices' compute budget is limited by battery life.

One of the most common tasks in egocentric vision is egocentric action recognition, which involves identifying and classifying the actions of a camera wearer based on the visual data captured from her point of view. Critically, the recording equipment is worn by the observer and it moves around with her. Hence, there is a far higher degree of changes in illumination, viewpoint and environment compared to a fixed third person camera. This variability leads to a notable drop in the performance of egocentric action recognition models when tested in conditions not seen during training. In general, this problem is referred to in the literature as *domain shift*, meaning that a model trained on a source labelled dataset cannot generalize well on an unseen dataset, called target, due to a discrepancy between their distributions. In egocentric vision, the domain shift is most commonly due to the so called "environmental bias" (Torralba and Efros, 2011). Given that video sequences are captured from a limited number of environments, training a model in one environment and deploying it in another leads to a performance decline due to intrinsic visual differences among them. Overcoming the environmental bias is essential to guarantee that models can operate reliably under the complex and unpredictable real-world conditions.

Humans have the ability to perceive the world around them through signals received from multiple sensory systems. Our perceptual experiences encompass visual, auditory, tactile, olfactory, and gustatory senses. Similarly, in egocentric vision, multi-modal information is crucial for understanding and disambiguating a user's intent or action. For instance, a video clip might display someone cutting tomatoes. While an activity recognition model based solely on video might not be able to categorize this from pure visual information alone, audio may provide a distinct sound that helps in recognizing the action (Morgado et al., 2021). The importance of multi-modal data is further amplified in egocentric vision due to the proximity of the device to where interactions occur. Audio data, in particular, becomes highly relevant

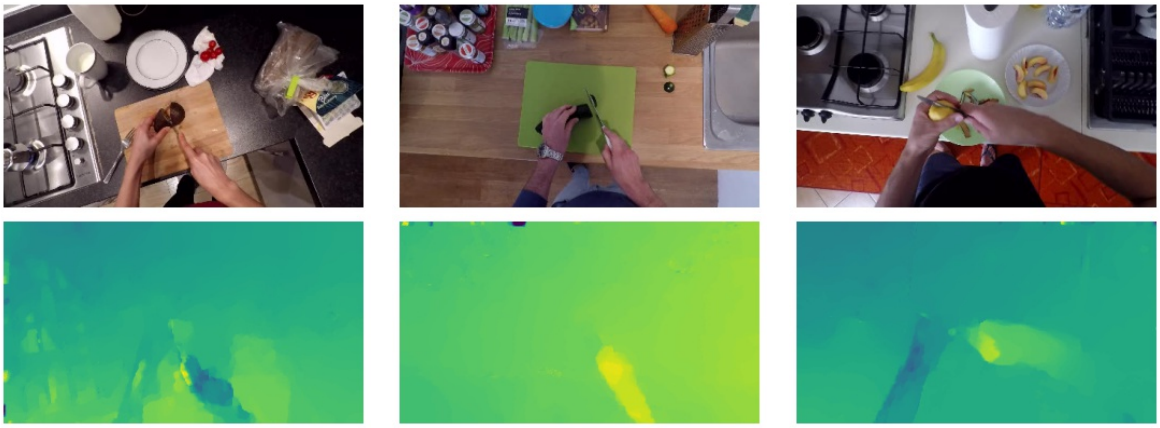


Fig. 1.1 **Egocentric vision across domains.** Data from RGB (top) and optical flow (bottom) across different environments. As it can be seen, optical flow information, focusing on domain-invariant information, is more robust to the domain shift.

as it captures interaction sounds that are crucial to accurately interpreting user's interactions. Moreover, different modalities are affected in different ways from the domain shift. For example, optical flow, being invariant to appearance (Sevilla-Lara et al., 2019), disregards domain-specific information such as the background, which can vary significantly across different environments (see Figure 1.1).

In this thesis, we propose to leverage multi-sensory information to tackle the complexities inherent in egocentric videos. In the first part of the thesis, we show how solving auxiliary tasks across different channels can improve action recognition generalization and adaptation to new domains. Specifically, we demonstrate how traditional modalities (RGB, audio, optical flow) can be integrated through an auxiliary loss which aligns their feature norms during training to improve performance on unseen environments. We then introduce a method based on video reconstruction through language for learning more robust features in the presence of a *scenario shift*, where the same action is performed as part of a different activity, and a *location shift*, where videos are recorded in different geographical locations. To study those domain shifts, we curated ARGO1M, the biggest dataset for action recognition generalization so far, including more than 1 million video clips.

In the second part of the thesis, we explore modalities not yet utilized in egocentric vision. We focus on the advantages of event-based data from event cameras, which excel in modeling motion information with less computational and power demands. With their high pixel bandwidth, dynamic range, low latency, and low power consumption, event cameras effectively address challenges like fast camera motion and background clutter typical of wearable devices (Gallego et al., 2020b). We introduce N-EPIC-Kitchens, the first dataset

for studying event-based data in egocentric vision, which extends the EPIC-KITCHENS dataset (Damen et al., 2018) to include event modality. Using this dataset, we benchmark the event modality against traditional ones, such as RGB and optical flow, to highlight its potential for egocentric action recognition.

Additionally, we explore the integration of 3D scene information as a new modality to overcome the limitations of incomplete scene views and limited field of view in 2D images. We introduce the task "Out of Sight, Not Out of Mind", which involves tracking object locations around the user over time, even when they are not visible, utilizing both frame-based images and 3D object positioning. This integration is particularly valuable as many modern wearable devices are equipped with SLAM (Simultaneous Localization and Mapping) technology, providing robust 3D positional data at no additional cost (Pan et al., 2023).

1.2 Research questions and motivations

Egocentric vision introduces several challenges not present in traditional third-person video understanding, primarily because of its unique way of capturing data. Key among these challenges is the so called "environmental bias", which hinders egocentric action recognition models' ability to generalize across different environments. This problem is most commonly known as *domain shift*. Tackling the challenge of domain shift is essential for enhancing the performance of learned models when applied to novel or unfamiliar environments. Many researchers in this field have addressed the problem of domain shift by reducing it to an Unsupervised Domain Adaptation (UDA) setting (Chen et al., 2019; Kim et al., 2021b; Munro and Damen, 2020a; Wei et al., 2022), where unlabeled samples from the target domain are available during training. However, the UDA scenario is not always practical, because the target domain might not be known in advance, or accessing target data at training time might be costly (or even impossible). An open research question is how to learn a representation capable of generalizing to any unseen domain, when it is not possible to access target data during training. This approach is most commonly referred to as the *Domain Generalization (DG)* setting. While this has been explored previously in image-based data for object classification tasks, we aim to investigate its application to the egocentric activity recognition task.

Moreover, it has been shown in the literature that certain modalities are inherently more robust to the domain shift (Munro and Damen, 2020a). For instance, cutting boards may differ in appearance (e.g., wooden vs. plastic), but optical flow overlooks this. Despite its

potential benefits, Multi-Modal Learning (MML) presents some challenges, such as learning how to summarize data while retaining their complementary information or understanding how to effectively combine information from multiple modalities for making predictions. Heterogeneity between modalities is another critical issue, as differences in their marginal distributions may prevent the model from learning equally from all of them. An open research question is how modalities that are different in nature, such as RGB and audio information, can be combined effectively. Drawing inspiration from recent works on self-supervised pretext tasks for learning representations from multi-modal content (Morgado et al., 2020; Munro and Damen, 2020a), we investigate how to solve auxiliary tasks across various video information channels in a manner that makes the solutions to such tasks consistent across channels and gains robustness from it.

Although optical flow is the most widely used modality along with RGB information, it demands high computational resources, limiting its application in real-time scenarios. Furthermore, it may not be ideal in a wearable context where saving battery and processing power is crucial. This opens a research question on whether event-based cameras, novel bio-inspired sensors that asynchronously capture pixel-level intensity changes as "events", might offer a solution. Due to their high pixel bandwidth, high dynamic range, low latency, and low power consumption, event cameras are well-suited for egocentric vision tasks, addressing challenges like fast camera motion and background clutter typical of wearable devices. Moreover, as they capture differential information, event sequences reveal more about scene dynamics than appearance, making them a valuable alternative to optical flow. We explore how those novel data behave in egocentric vision, assessing their effectiveness in enhancing action recognition accuracy and computational efficiency.

Finally, a major challenge in egocentric vision is the camera's limited field of view, which captures only a portion of the broader scene, thereby significantly constraining a comprehensive understanding of the environment. This challenge is compounded by the dynamic nature of human interaction with their surroundings, as objects frequently move in and out of the camera's field of view. On the other side, recent advances in 3D scene reconstruction (Tschernezki et al., 2024) unlock the possibility to represent in 3D coordinates the environments in which the videos have been recorded. This introduces research questions on how to merge 3D scene information (complete observation) with partial observations from RGB images to enhance our perception of dynamic environments. We analyze the impact of 3D information about objects the user interacts with in the scene to enable the tracking of multiple dynamic objects over time, providing a continuous understanding of the location of all objects, even when they are not visible in the field of view.

To summarize, this thesis aims to answer a number of questions that have yet to be answered. In particular, *how can we leverage the benefits of multi-modal learning to mitigate domain differences and enhance the robustness of egocentric action recognition models on unseen domains, particularly when we lack access to data from the test distribution? What are the key challenges in integrating event-based cameras into traditional egocentric vision models, and could this modality prove to be truly beneficial in enhancing the models' performance and adaptability? Can comprehensive 3D information about the scene where videos are recorded be integrated with partial 2D images from the camera's limited field of view to achieve a complete understanding of the user's surroundings?*

1.3 Outline and main contributions

In this thesis, we propose to address the research questions outlined above through the development of two multi-modal Domain Generalization (DG) frameworks for egocentric action recognition. The first one focuses on domain generalization across various environments by aligning the feature norms of multiple modalities – RGB, optical flow, and audio – during training, as detailed in Chapter 3. The second one aims at enhancing action recognition generalization across different scenarios and locations using textual information. To investigate this problem, we introduce ARGO1M, the largest dataset created for action recognition generalization. Subsequently, we propose a DG method based on a visual-language reconstruction task designed to effectively address domain shifts (Chapter 4).

We then move our investigation on the introduction and analysis of new modalities within egocentric vision. The use of event data is explored in Chapter 5. We introduce a new event-based egocentric vision dataset obtained through event data simulation, and use it benchmark event-data w.r.t. traditional modalities. We then show how domain adaptation techniques can be employed to ensure strong performance on real event-based data, when training occurs exclusively on simulated samples. Finally, we demonstrate how information about objects' location in the 3D scene where videos have been recorded can mitigate problems associated with the limited field of view in egocentric cameras (Chapter 6). The thesis is structured as follows:

- Chapter 2 begins with a general overview of the potential applications and benefits of egocentric vision. It then continues with a description of existing models for egocentric action recognition, highlighting both seminal and state-of-the-art works, and an overview of 3D egocentric scene understanding tasks. We then present a

detailed overview of existing Domain Generalization (DG) and Unsupervised Domain Adaptation (UDA) techniques, focusing specifically on those employed in this thesis. Finally, we discuss the functioning of event-based cameras, outlining their advantages over traditional vision devices.

The chapter contains part of the work in (Plizzari et al., 2023a), published at the International Journal of Computer Vision in 2024 (IJCV).

- Chapter 3 introduces a multi-modal framework for DG in egocentric action recognition. This chapter discusses the "imbalance" problem that arises when training multi-modal networks, which often leads to the network favoring one modality over others, thereby diminishing its generalization capabilities. To tackle this issue, we propose a novel multi-modal loss designed to progressively align the relative feature norms of multiple modalities (RGB, audio, and optical flow) during training. Our results demonstrate that rebalancing the contribution of these modalities during training leads to improved generalization performance. Additionally, we extend this method to operate under the UDA setting, utilizing unlabeled target data, where we also confirm the effectiveness of our approach in this context.

  The chapter led to the publication of (Planamente et al., 2022b) at the Winter Conference of Computer Vision in 2022 (WACV22) and of (Planamente et al., 2024) at International Journal of Computer Vision in 2024 (IJCV). The proposed method also achieved the third place in the EPIC-Kitchens Unsupervised Domain Adaptation Challenge at the Computer Vision and Pattern Recognition conference in 2021.

- Chapter 4 addresses the challenge of domain action recognition generalization across various scenarios and locations. To support this research, we have curated the Action Recognition Generalization dataset (ARGO1M). We then present a domain generalization approach that incorporates Cross-Instance Reconstruction along with video-text pairing. This strategy aims at learning representations that are robust and generalizable across diverse conditions by enhancing the model's capacity to comprehend and adapt to new, unseen environments. Our results demonstrate that textual information is instrumental in guiding the development of representations that are more robust on the challenging ARGO1M dataset.

  The chapter led the the publication of (Plizzari et al., 2023b) at the International Conference of Computer Vision in 2023 (ICCV23).

- Chapter 5 delves into the utilization of event data in egocentric vision. Initially, we introduce N-EPIC-Kitchens, the first event-based egocentric action recognition dataset,

which enables the exploration of event data in this domain. We then propose two event-based approaches, E^2(GO) and E^2(GO)MO, designed to exploit the motion information captured by event data for egocentric action recognition. Our findings indicate that event-based data achieves performance on par with RGB in seen environments and even surpasses RGB in unseen ones. In the second part of the chapter, we demonstrate how unsupervised domain adaptation methods can effectively bridge the simulated-to-real (Sim-to-Real) gap for event cameras by aligning the feature distributions between a simulated source domain and the real target domain.

The chapter led the the publication of three works. The first one (Plizzari et al., 2022) is published at the Computer Vision and Pattern Recognition Conference in 2022 (CVPR22). The second one (Cannici et al., 2021) is published at the International Workshop on Event-based Vision and Smart Cameras, held at the 2021 Conference on Computer Vision and Pattern Recognition. The last one (Planamente et al., 2021) is published at the Robotics and Automation Letter journal in 2022 (RA-L) and presented at the 2021 International Conference on Intelligent Robots and Systems (IROS21).

- Chapter 6 explores the integration of 3D information about object locations as a new modality, combined with partial 2D observations from egocentric videos, to achieve a comprehensive understanding of the environment. This approach helps overcome the limitations imposed by the camera's narrow field of view. We introduce the "Out of Sight, Not Out of Mind" task, which involves tracking multiple objects over time, even when they temporarily leave the field of view. Our findings demonstrate that 3D information plays a crucial role in accurately maintaining continuity and awareness of where objects are, enhancing the overall effectiveness of egocentric vision systems.

  This chapter is contained in a preprint article (Plizzari et al., 2024)

- Chapter 7 concludes the thesis with a summary of the work presented and outlines potential future directions.

1.4 Publications

In the following section, the main articles in this thesis are listed[2]:

1. Planamente, M., Plizzari, C., Peirone, S. A., Caputo, B., & Bottino, A. (2024). Relative Norm Alignment for Tackling Domain Shift in Deep Multi-modal Classification. *International Journal of Computer Vision, 1-21.*
   Online Resources: [Paper]

2. Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen, D. & Tommasi, T. (2024). An outlook into the future of egocentric vision. *International Journal of Computer Vision.*
   Online Resources: [Paper]

3. Plizzari, C., Goel, S., Perrett, T., Chalk, J., Kanazawa, A., Damen, D. (2024). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. *Preprint.*
   Online Resources: [Paper], [Project page]

4. Plizzari, C., Perrett, T., Caputo, B., & Damen, D. (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13656-13666).*
   Online Resources: [Paper], [Project page]

5. Plizzari*, C., Planamente*, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., & Caputo, B. (2022). E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 19935-19947).*
   Online Resources: [Paper], [Project page]

6. Planamente*, M., Plizzari*, C., Alberti, E., & Caputo, B. (2022). Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1807-1818).*
   Online Resources: [Paper]

7. Plizzari*, C., Planamente*, M., Alberti, E., Caputo, B., PoliTO-IIT Submission to the EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action

[2] $*$ indicates equal contribution

Recognition.
Third Place at the EPIC-Kitchens Unsupervised Domain Adaptation Challenge at CVPR 2021. (technical report)
Online Resources: [Paper]

8. Planamente*, M., Plizzari*, C., Cannici*, M., Ciccone, M., Strada, F., Bottino, A. & Caputo, B. (2021). Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *IEEE Robotics and Automation Letters, 6(4), 6616-6623.*
Online Resources: [Paper]

9. Cannici*, M., Plizzari*, C., Planamente*, M., Ciccone, M., Bottino, A., Caputo, B., and Matteucci, M. (2021). N-ROD: A Neuromorphic Dataset for Synthetic-to-Real Domain Adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 1342-1347).*
Online Resources: [Paper], [Project page]

The following is a list of the additional research contributions not expliciteply covered in the thesis:

1. Nasirimajd, A., Plizzari, C., Peirone, S., Ciccone, M., Averta, G., Caputo, B., (2024). *Domain Generalization using Action Sequences for Egocentric Action Recognition. Under submission at the IEEE Robotics and Automation Letters journal*

2. Nasirimajd, A., Peirone, S., Plizzari, C., Caputo, B., (2023). EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge: Mixed Sequences Prediction.
Second Place at the EPIC-Kitchens Unsupervised Domain Adaptation Challenge at CVPR 2023. (technical report)
Online Resources: [Paper]

3. Neubert, J., Planamente, M., Plizzari, C., & Caputo, B. (2023). LCMV: Lightweight Classification Module for Video Domain Adaptation. In *International Conference on Image Analysis and Processing (pp. 270-282). Cham: Springer Nature Switzerland.*
Online Resources: [Paper]

4. Plananamente*, M., Plizzari*, C., & Caputo, B. (2022). Test-time adaptation for egocentric action recognition. In *International Conference on Image Analysis and Processing (pp. 206-218). Cham: Springer International Publishing.*
Online Resources: [Paper]

5. Plizzari, C., Cannici, M., & Matteucci, M. (2021). Spatial temporal transformer network for skeleton-based action recognition. In *Pattern recognition. ICPR international workshops and challenges: virtual event, January 10–15, 2021, Proceedings, Part III (pp. 694-701). Springer International Publishing.*
   Online Resources: [Paper]

6. Plizzari, C., Cannici, M., & Matteucci, M. (2021). *Skeleton-based action recognition via spatial and temporal transformer networks. Computer Vision and Image Understanding, 208, 103219.*
   Online Resources: [Paper]

Chapter 2

Preliminaries

This chapter provides an overview of existing tasks in egocentric vision, and discusses potential applications of egocentric vision (Section 2.1). For a broader overview of egocentric vision tasks, we direct the reader to the comprehensive survey in (Plizzari et al., 2023a). Additionally, the chapter provides an overview of cross-domain challenges in Section 2.2, and discusses methodologies developed for learning across domains. Finally, it offers a detailed description of the working principles of event-based cameras and explores their integration within deep learning architectures for computer vision, as detailed in Section 2.3.

2.1 Egocentric Vision

We offer an overview of existing tasks in egocentric vision, as well as a vision for the future, through character-based stories and associated visuals (Section 2.1.1). In each narrative, we explore various research tasks associated with egocentric vision applications. This thesis specifically delves into two primary research tasks: *action recognition* and *3D scene understanding*. In Section 2.1.2 we describe existing seminal and state-of-the-art approaches to action recognition, and in Section 2.1.3 we illustrate the advancements in 3D understanding. For both tasks, we discuss datasets tailored to these objectives, alongside their limitations and potential future applications.

2.1.1 An Outlook into Egocentric Vision

We present two use cases, each rooted in specific locations or professions. For each use case, we summarize the relevant existing technologies before presenting futuristic scenarios through short, character-driven narratives, enhanced with illustrations drawn by an artist to spark the readers' imagination. The main figures in these stories utilize a wearable device named *EgoAI*, which offers in-situ multi-modal sensing from the user's perspective, providing personalized, ego-centric assistance. We then explore more in-depth the connection between these use cases and existing research tasks.

EGO-Worker

Current large-scale workshops and factories are increasingly incorporating vision-based systems, yet these systems predominantly depend on stationary cameras. To cover various areas, these cameras must be installed throughout the facility, but they offer only a limited viewpoint, thus limiting their effectiveness. The process of training and supervising workers typically relies on prerecorded materials or direct guidance from more experienced colleagues. However, this method often results in a loss of expertise when an employee leaves for another job. Moreover, the feedback provided to employees regarding their performance usually comes from heuristic evaluations, either automated or manual, which may not accurately reflect their true performance. Additionally, this feedback is often not effectively linked with training or guidance on how to enhance their skills. Although technology's role in ensuring worker safety is growing, it has not met the expectations that come with technological advancements, which tend to focus more on increasing productivity than improving safety measures. EgoAI will bridge these gaps, aiming to improve workplace safety and comfort, offering a more cohesive, effective, and comprehensive approach to worker training, monitoring, and feedback.

Every morning, Marco starts his shift with a routine check in front of the mirror, allowing EgoAI to confirm that he's correctly wearing his Personal Protective Equipment (PPE) to ensure his safety. Following this verification, Marco inquires with EgoAI about his assigned location within the factory for the day. EgoAI accurately pinpoints Marco's position and guides him to his workstation, skillfully avoiding areas with overhead hazards and paths designated for vehicle movement. Marco has complete faith in EgoAI's navigation, recalling a time when it efficiently directed him to the nearest fire extinguisher to prevent a fire from spreading.

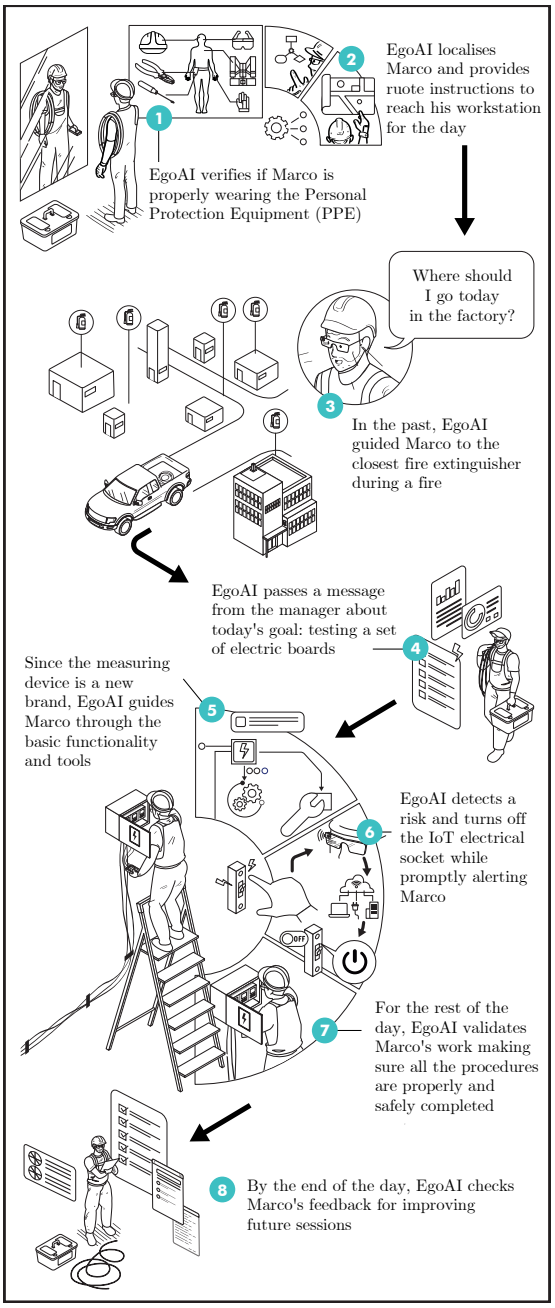


Fig. 2.1 **EGO-Worker**. Illustration of the story from Sec 2.1.1. EgoAI assists Marco from the beginning to the end of his day. 1 Safety Compliance Assessment. 2 5 Localization and Navigation. 4 Messaging. 5 Hand-Object Interaction. 6 Action Anticipation. 7 Skill Assessment. 8 Visual Question Answering, 8 Summarization.

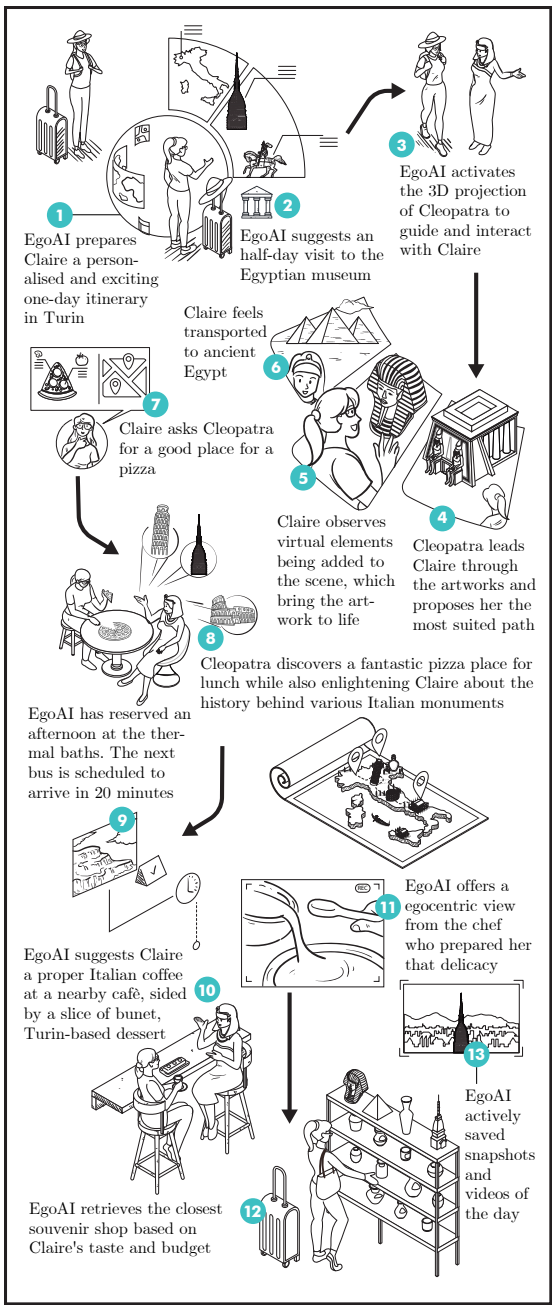


Fig. 2.2 **EGO-Tourist**. Illustration of the story from Sec 2.1.1. EgoAI accompanies Claire throughout her itinerary in Turin. 1 2 8 9 10 11 Recommendation and Personalization. 2 3 4 5 6 3D Scene Understanding. 5 Gaze Prediction. 3 4 8 12 Localization and Navigation. 7 Messaging. 8 Visual Question Answering. 11 Action Recognition and Retrieval. 13 Summarization.

Upon arriving at his station, Marco receives a directive from his manager through EgoAI to test a series of electrical boards. With the measuring tool being a new model, EgoAI walks Marco through its essential functions to ensure the tests on the boards are conducted accurately. However, Marco becomes momentarily distracted and nearly attempts to probe an electrical board while it's powered on. EgoAI immediately identifies the danger and deactivates the IoT electrical socket connected to the board, simultaneously warning Marco.

Throughout the day, EgoAI oversees Marco's activities, ensuring that all tasks are performed correctly and safely. It provides assistance whenever Marco has questions, monitors his stress levels, and prompts him to take necessary breaks.

As the day concludes, EgoAI expresses gratitude to Marco for his diligence, especially with the new procedures, and seeks his input for improving training methods. EgoAI then automatically incorporates Marco's feedback and suggestions into the development of future training sessions and plans, continuously enhancing the workplace environment and safety protocols.

EGO-Tourist

Today, travelling for tourism and vacations has seen a remarkable increase, more than doubling in frequency in the past 20 years[1]. In recent years, there has been a growing fusion between technology and art, spanning from the ancient to the contemporary. This integration has enhanced the accessibility and interactive potential of art through technological advancements. Digital audio guides and virtual tours are becoming the norm in museums and tourist attractions, playing a vital role in engaging visitors and enriching their experience. However, despite these advancements, the personal touch in the visitor experience is often missing, requiring active participation from the users to truly benefit. EgoAI steps in to bridge these gaps, transforming travel into an enjoyable and interactive adventure by providing personalized experiences tailored to each user's interests and preferences.

Arriving in Turin as the final destination of her Italian vacation, Claire is eager to explore the city but lacks detailed knowledge about it. Fortunately, EgoAI is well attuned to Claire's preferences and has crafted a personalized and thrilling one-day itinerary just for her. Knowing her keen interest in museums, EgoAI allocates half the day for a visit to the renowned local Egyptian Museum. During her exploration, EgoAI enhances the experience by activating a 3D projection of Cleopatra to serve as Claire's guide and interactive companion, leading her through the museum and recommending the most intriguing path.

[1] https://ourworldindata.org/tourism

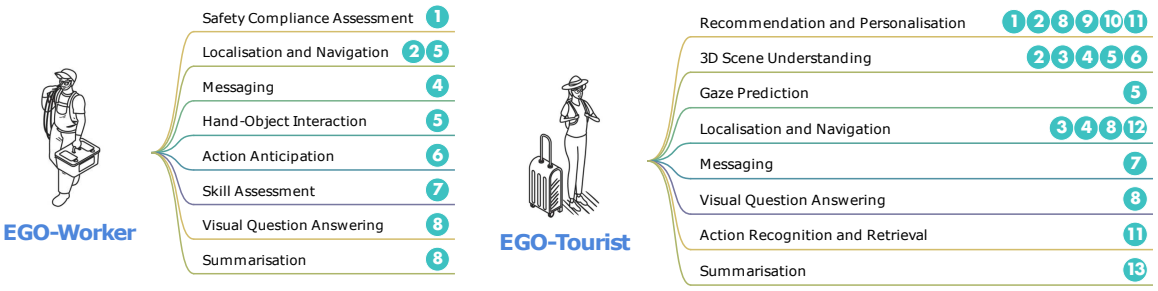


Fig. 2.3 **Connections between narratives and the research tasks.** For each of the use cases presented in Section 2.1.1, we show the corresponding research tasks, along with the specific part of the story where the tasks are occurring, indicated by the numbers corresponding to those representing sub-stories in Figures 2.1 and Figure 2.2 respectively.

While engaging with Cleopatra about a sarcophagus, Claire witnesses virtual elements being integrated into her surroundings, bringing ancient artworks to life. This immersive experience transports her back to the times of ancient Egypt, allowing her to interact with and understand the historical artifacts in their intended context.

After the museum visit, Claire decides to keep Cleopatra as her augmented reality (AR) guide for lunch, seeking recommendations for a great pizza spot. Over lunch, she continues her conversation with Cleopatra, gaining deeper insights into the Italian monuments she visited earlier in her trip, enhancing her understanding of their historical significance.

With the afternoon planned for relaxation at the thermal baths, and the next bus scheduled in 20 minutes, EgoAI suggests Claire enjoy an authentic Italian coffee accompanied by a slice of bunet, a famous dessert from Turin. Curious about the dessert's recipe, EgoAI provides Claire with a first-person tutorial from the chef who prepared it.

Following her time at the thermal baths, EgoAI inquires if Claire wishes to purchase souvenirs for her family. It then locates the nearest souvenir shop that matches her relatives' tastes and her budget.

Throughout her day in Turin, Claire was fully immersed in her experiences, free from the concern of documenting the moments herself. EgoAI proactively captured significant snapshots and videos of her favorite moments, ensuring that her memories of the trip are preserved without her needing to lift a finger.

From Narratives to Research Tasks

Various research tasks can be identified in the above character-based narratives/stories. In this

section, we link the above narratives to research challenges as recognized by the academic community. Additionally, we assess if these challenges can be addressed with current wearable technologies or if there is a need for newer, more sophisticated devices to surpass the constraints of those presently in the market. The relationships between our use cases and these research tasks are depicted in Figure 2.3.

For tasks that utilize augmented reality (AR) technology, a comprehensive understanding of 3D scenes becomes essential. This requirement is highlighted in scenarios such as EGO-Tourists' immersive museum visits. The envisioned AR technology is further enhanced with directional audio synthesis, adding auditory feedback to increase the realism of the augmented surroundings.

Navigating through a 3D environment necessitates the tasks of localization and navigation, which are pivotal, regardless of the space's constraints. This requirement is evident in the factory scenario presented in EGO-Worker. The capability of contemporary egocentric devices to interpret 3D spaces is progressively advancing, thanks to the incorporation of more recent cameras (e.g., Microsoft HoloLens 2[2], Xreal Light[3], Magic Leap 2[4], Project Aria Glasses[5]). These devices are capable of scanning the surrounding area to construct a 3D model of the static environment, thereby facilitating the localization of the user and simplifying navigation. However, dynamic scenes and outdoor environments continue to pose significant challenges to these systems, making the realistic integration of 3D scene understanding into practical applications an ongoing area of research.

Within the scene, the process of comprehensively understanding actions is executed through tasks such as action recognition, which experiences a significant shift as the perspective changes from third-person to first-person views. In scenarios like EGO-Worker, the device plays a crucial role in validating the user's actions in a work environment. Notably, the aspect of action anticipation stands out, as the device is equipped to quickly identify and prevent potentially dangerous situations before they occur. Currently, the market lacks smart glasses capable of robustly recognizing human actions in real-time.

EgoAI is enhanced with *gaze prediction* technology, allowing it to monitor the user's eye movements and smoothly align with the user's gaze towards objects. This functionality is evident in scenarios like EGO-Tourist, where users interact with museum artifacts. While gaze tracking technology has reached a level of reliability, it still necessitates an initial eye calibration and may experience accuracy drift over time. Wearable devices such as the

[2] https://www.microsoft.com/en-us/hololens
[3] https://www.xreal.com/light/
[4] https://www.magicleap.com/magic-leap-2
[5] https://about.meta.com/realitylabs/projectaria/

Microsoft Hololens2, Magic Leap 2, Project Aria Glasses, and Apple Vision Pro have already incorporated this feature[6].

Hand-pose estimation and *hand-object interactions* are crucial for the effectiveness of *EgoAI*. In EGO-Worker, *EgoAI* aids in the operation of unfamiliar measuring tools, showcasing its ability to facilitate direct interaction with new equipment.

The success of the envisioned *EgoAI* device will also hinge on its ability to handle a variety of supplementary tasks. The feature of *messaging* is a recurring theme in the narratives. In EGO-Worker, messages from the manager about daily tasks are receive.

The convenience of hands-free operation is further augmented by the implementation of voice commands, facilitating effortless interaction. This is evident in EGO-Tourist, where the tourist requests further details about an artwork through voice queries. Modern wearable glasses often incorporate voice assistants like Microsoft's Cortana[7], Apple's Siri[8], or Google Assistant[9]. These assistants enhance user interaction by enabling them to open apps, capture photos, send messages, and much more, significantly enriching the user experience.

Another crucial function of *EgoAI* is Safety Compliance Verification. In EGO-Worker, it verifies whether the worker is properly outfitted with Personal Protection Equipment (PPE) using advanced recognition and identification methods.

In this thesis, we investigate the task of *action recognition* across multiple modalities (audio and optical flow - Chapter 3, language - Chapter 4) and domains. In Chapter 5 and Chapter 6 we introduce event-based data and 3D information as new modalities for effective video understanding. In the following, we delve into existing approaches for the task of *action recognition* (Section 2.1.2) and *3D scene understanding* (Section 2.1.3), which are the most relevant for this work.

2.1.2 Egocentric Action Recognition

The goal of egocentric action recognition is to recognize actions from a first-person perspective. In contrast to third-person action recognition, where the camera observes the scene from an external viewpoint, egocentric vision involves processing visual data captured from the point of view of the participant. This means the camera is typically mounted on the person's body, often on the head or chest, capturing what the wearer sees. This is a relatively new

[6] https://www.apple.com/apple-vision-pro/
[7] https://www.microsoft.com/en-us/cortana
[8] https://www.apple.com/siri/
[9] https://assistant.google.com/

task which has already found many applications in ambient assisted living (Meditskos et al., 2018; Nakazawa and Honda, 2019; Zhan et al., 2014), augmented reality (AR) and virtual reality (VR) technologies (Liang et al., 2015; Taylor et al., 2020), and social interaction analysis (Aghaei et al., 2016; Alletto et al., 2014; Fathi et al., 2012a; Ryoo and Matthies, 2013). Recognizing actions from an egocentric viewpoint introduces a greater level of complexity due to the camera's movement, which is dynamic and often unpredictable, unlike the static perspective offered by fixed, external cameras. An additional challenge is the camera wearer's presence in the visual field, which causes occlusions and an only partial visibility of the action performed. To overcome these obstacles, one strategy involves utilizing additional modalities alongside visual data. For instance, audio signals, the wearer's gaze direction, and motion patterns captured through optical flow can significantly enhance action recognition. However, integrating multiple data types can be resource-intensive. Consequently, recent progress in this area has been directed towards developing energy-efficient architectures that also excel in interpreting complex actions at a higher level.

In the following, we describe the main methods architectures for addressing the action recognition task. Given that egocentric action recognition models draw upon standard third-person action recognition frameworks without being specifically tailored for a first-person viewpoint, we also examine action recognition models developed for third-person vision. We then introduce datasets for egocentric action recognition and discuss current limitations and future works for this task.

Action recognition methods and architectures

Early works leveraged the egocentric perspective to improve action recognition for robots (Johnson and Demiris, 2005) and humans (Surie et al., 2007). In the pre-deep learning era, approaches for egocentric action recognition mainly included the use of descriptors like Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005), and Histogram of Optical Flow (HOF) (Wang and Snoussi, 2013) to extract features. Among those, (Spriggs et al., 2009) explored action recognition for egocentric vision accompanied with Inertial Measurement Units (IMUs), and (Kitani et al., 2011) was the first to tackle action recognition from egocentric sports videos in an unsupervised manner. (Kitani et al., 2011) used motion-based histograms recovered from the optical flow of the scene to learn the action categories performed by the wearer. The research field received significant attention following the release of a dataset featuring activities of daily living (ADL) (Pirsiavash and Ramanan, 2012), notably for its comprehensive annotations covering activities, object trajectories, hand positions, and interaction incidents.

(Fathi et al., 2012b) pioneered in demonstrating the importance of gaze by introducing a probabilistic generative model that concurrently identifies daily activities and anticipates gaze points in egocentric video footage. (Li et al., 2015) introduced an approach that integrates hand posture, head movement, and gaze orientation features with those of motion and objects to enhance the analysis of egocentric videos.

In recent years, deep learning has significantly reduced the need for manually extracting features. The 2D Convolutional Neural Networks (CNNs) (Kazakos et al., 2019b; Poleg et al., 2016; Ryoo et al., 2015; Singh et al., 2016), originally designed for image analysis, were adopted for video processing, treating each video frame as an individual image. Approaches employing recurrent neural networks like Long Short-Term Memory (LSTM) (Cao et al., 2017; Verma et al., 2018) and Convolutional Long Short-Term Memory (ConvLSTM) (Sudhakaran and Lanz, 2017, 2018) have been developed to more effectively capture temporal dynamics. To further modelling motion dynamics in videos, two-stream networks were introduced, simultaneously processing the spatial and temporal streams of the video (Kazakos et al., 2019b; Tang et al., 2017). 3D Convolutional Neural Networks (3D CNNs) (Carreira and Zisserman, 2017; Feichtenhofer et al., 2019; Hara et al., 2017; Ji et al., 2012; Tran et al., 2015, 2018) were proposed to inherently capture both spatial and temporal features by extending convolutions into the temporal domain. The advent of the Transformer architecture (Vaswani et al., 2017) has inspired a series of studies utilizing transformers as a core framework for video processing (Arnab et al., 2021; Patrick et al., 2021). These efforts expand upon the Vision Transformer (Dosovitskiy et al., 2020), adapting it to handle sequences of frames.

In the following, we describe 2D CNN-based methods, as well as 3D CNN-based ones, and recent Transformer-based architectures. An illustration of the distinction between 2D CNN-based methods and 3D CNN-based ones is shown in Figure 2.4.

2D CNN-based methods. Convolutional Neural Networks (CNNs), particularly 2D CNNs, are widely employed in image classification tasks due to their efficiency in learning hierarchical visual patterns. The adaptation of 2D-CNNs for video action recognition requires the integration of temporal dynamics with the spatial feature extraction capabilities intrinsic to 2D-CNNs. A straightforward method involves sliding the convolutional kernel, having dimensions $k \times k$, across the complete set of video frames. An advanced approach to incorporate the temporal aspect is represented by the Temporal Segment Network (TSN) framework (Wang et al., 2016), a seminal method based on 2D-CNNs. TSN strives to comprehend both the spatial attributes of individual frames, and the sequential arrangement of these frames. It divides the video into multiple segments, retrieves a brief snippet from each segment, and

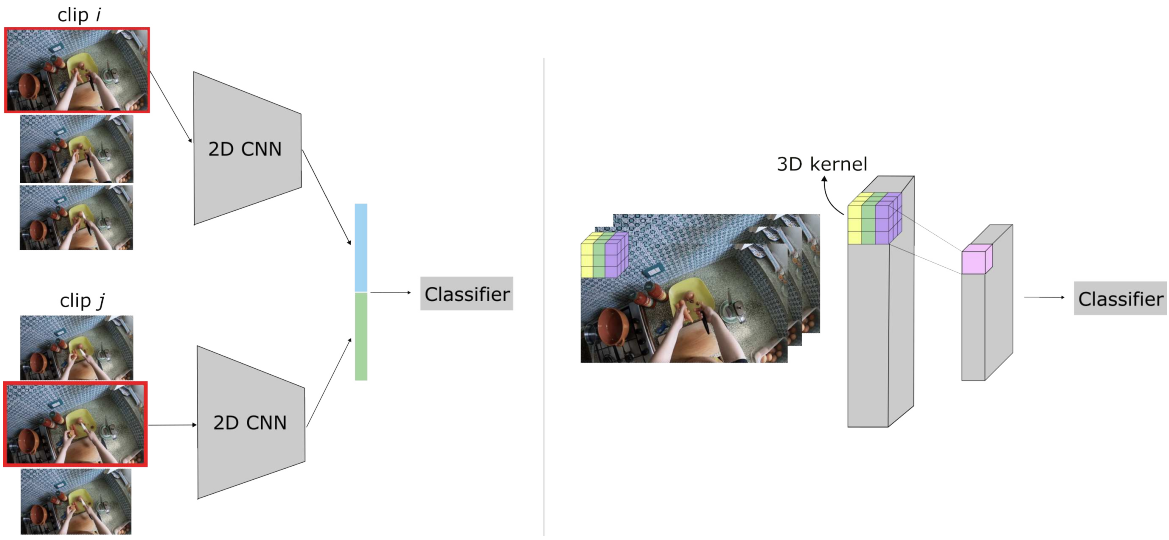


Fig. 2.4 **2D CNN-based *vs* 3D CNN-based methods.** On the left, two different video clips, *i* and *j*, are processed independently through a 2D CNN before classification. On the right, a 3D CNN processes a stack of consecutive frames capturing both spatial and temporal information, which is then fed into the classifier.

employs these snippets as inputs to the network. Subsequently, TSN combines the classification output from each segment to formulate a video-level inference. The TSN methodology employs a sparse sampling strategy in the temporal dimension to efficiently capture the dynamics of motion. Several works have been designed to extend this approach (Wang et al., 2017b; Zhou et al., 2018). Temporal Relation Network (TRN) (Zhou et al., 2018) has been introduced to explicitly model the temporal relations among video frames. TRN has a hierarchical structure endowing it with heightened sensitivity towards the sequential ordering of frames, thereby equipping it with the capability to process and interpret more intricate actions that necessitate a deeper understanding of temporal dynamics. (Kazakos et al., 2019b) introduced an end-to-end trainable mid-level fusion model known as the Temporal Binding Network (TBN), which is built upon a 2D convolutional network. This model is designed to asynchronously integrate audio, RGB, and optical flow information across various temporal windows.

2D CNNs are good in extracting spatial representations, yet their efficacy in temporal dimension encoding is not as robust. In response, several approaches (Du et al., 2017; Meng et al., 2020; Perrett and Damen, 2019; Sudhakaran and Lanz, 2017, 2018; Sudhakaran et al., 2019; Sun et al., 2017) have incorporated Recurrent Neural Network (RNN) architectures, notably Long-Short Term Memory (LSTM) (Memory, 2010), to model the long-range temporal context of video sequences, building upon the spatial features obtained through CNNs. An innovative recurrent module has been introduced by (Sudhakaran et al., 2019),

enhancing LSTM with an inherent spatial attention mechanism and a modified output gate. This facilitates focusing on relevant spatial regions and ensures seamless tracking of attention throughout the video frames. However, action recognition methods utilizing RNNs often overemphasise temporal aspects, potentially compromising their ability to extract distinctive spatial features and leading to sub-optimal performance. Moreover, RNNs are prone to vanishing and exploding gradient problems, especially when processing long sequences, which can affect their efficiency in learning temporal dynamics.

Two-stream methods have been introduced to capture both spatial and temporal information while effectively addressing long-term dependencies. They leverage both RGB and optical flow information to capture appearance and motion cues respectively. (Simonyan and Zisserman, 2014) proposed an innovative two-stream model utilizing two streams for video analysis. The first, a spatial stream, applies a 2D CNN, like AlexNet (Krizhevsky et al., 2012) or VGGNet (Simonyan and Zisserman, 2014), to single video frames, focusing on the extraction of spatial attributes such as the appearance of objects, their shapes, and the surrounding context. In contrast, the second stream, the temporal stream, processes optical flow frames that depict movement between successive frames. Techniques such as the Farneback method (Farnebäck, 2003) or the TV-L1 algorithm (Zach et al., 2007) are employed to compute the optical flow. This stream leverages a CNN to distill features representing motion dynamics. The integration of the spatial and temporal features is achieved through two primary fusion approaches, with *late fusion* combining the softmax probabilities from both streams to output the final classifications, and *early fusion* directly combining the features from both streams. Several works (Feichtenhofer et al., 2017; Girdhar et al., 2017; Zong et al., 2021) have further developed the two-stream architecture to enhance the comprehension of extensive video content. The main challenge in the two-stream model is the optical flow computation, which is traditionally resource-intensive. To alleviate this, (Zhang et al., 2016) suggested substituting motion vectors extracted from compressed video data in place of optical flow. Although this significantly speeds up the computation, it comes with a compromise, as motion vectors are generally less detailed and more susceptible to noise, leading to a degradation in recognition accuracy.

3D CNN-based methods. Several studies have adopted 3D Convolutional Neural Networks (CNNs) for feature extraction. Unlike 2D CNNs, the convolutional layers in a 3D CNN extend their kernels across height, width, and depth – the latter representing time in video analysis or the z-axis in 3D volumetric data. This architectural distinction enables 3D CNNs to simultaneously process spatial and temporal information, making them particularly effective for video analysis tasks. This section explores various methodologies that employ

3D CNNs. A considerable number of works (Carreira and Zisserman, 2017; Feichtenhofer et al., 2019; Hara et al., 2017; Ji et al., 2012; Tran et al., 2015, 2018) have expanded upon the capabilities of 2D CNNs by integrating 3D CNNs to capture spatial-temporal features, thereby achieving improved video comprehension.

(Tran et al., 2015) introduced 3D ConvNets, capturing both spatial and temporal aspects of video data by performing convolution and pooling operations spatio-temporally. The 3D ResNet (R3D) (Hara et al., 2017) adapts the robust ResNet (He et al., 2016) framework to a 3D context for video action recognition, offering a straightforward yet potent structure for direct spatio-temporal feature extraction from videos. Inflated 3D ConvNets (I3D) (Carreira and Zisserman, 2017) is based on 2D ConvNet inflation: filters and pooling kernels of very deep image classification ConvNets are expanded into 3D, making it possible to learn spatio-temporal feature extractors from video while leveraging successful ImageNet architecture designs and even their parameters expand filters and pooling layers to 3D, optimizing the use of C3D (Tran et al., 2015) for more efficient video analysis. The R(2+1)D model (Tran et al., 2018) segments traditional 3D convolutions into a sequence of 2D spatial convolutions followed by 1D temporal convolutions, based on the R3D framework (Hara et al., 2017). This approach enables the network to separately learn spatial and temporal features, enhancing efficiency. SlowFast (Feichtenhofer et al., 2019) introduces a dual-pathway architecture: a "slow" path for capturing spatial semantics at a lower frame rate, and a "fast" path for high-temporal-resolution motion detection. Either C3D (Tran et al., 2015) or R3D (Hara et al., 2017) can be integrated within these pathways to extract spatio-temporal features effectively. Conversely, several studies (Lin et al., 2019; Sudhakaran et al., 2020; Wang et al., 2021a,b) have employed 2D CNNs coupled with advanced temporal modules to mitigate the computational demands of 3D CNNs. Temporal Shift Module (TSM) (Lin et al., 2019) introduces an efficient yet powerful mechanism, the shift operation, which facilitates the analysis of temporal sequences at no extra computational cost. This operation redistributes a portion of the channels across the temporal axis, thereby enhancing inter-frame communication. Unlike TSM's (Lin et al., 2019) parameter-free shift, the Gate-Shift Network (GSM) (Sudhakaran et al., 2020) addresses the varying degrees of motion dynamics within and across different action categories by adaptively modulating feature maps along the temporal dimension.

Traditional 3D CNN approaches to video understanding typically employ window-based convolutions that focus on short spatio-temporal segments, which constrains their capacity to extract long-term dependencies. Transformer-based models have emerged as a prominent solution in action recognition, thanks to their ability to directly process entire video sequences. By utilizing a scalable self-attention mechanism (Vaswani et al., 2017), transformers excel at

comprehending extensive spatio-temporal correlations, marking a significant advancement in the field. The next section will review transformer-based methods.

Transformer-based methods. Transformer-based approaches have significantly revolutionized the domain of action recognition, offering an advanced framework for comprehending long-range dependencies and global context within video sequences. Utilizing self-attention mechanisms and positional encoding, transformers have been introduced to extract spatio-temporal correlations, showcasing state-of-the-art performance across a range of tasks.

The Vision Transformer (ViT) (Dosovitskiy et al., 2020), marking the initial application of Transformer self-attention (Vaswani et al., 2017) in the realm of computer vision, introduces a novel approach by representing images as sequences of patches. This technique uses the self-attention mechanism for capturing the global context and inter-patch dependencies. Specifically, an input image is segmented into fixed-size patches, each of which is linearly transformed into a sequence of tokens. These tokens are subsequently processed by a traditional transformer encoder. The self-attention mechanism facilitates each token's interaction with others, enabling the modeling of long dependencies. Spatial information regarding the positions of patches within the image is incorporated through positional encodings added to the token embeddings. For image classification tasks, ViT introduces a classification token at the beginning of the sequence, which is processed by the transformer encoder. The classification token's output is fed to a classification head, such as a fully connected layer, to generate final class predictions. ViT has demonstrated impressive results on various image classification benchmarks, competing with traditional Convolutional Neural Networks (CNNs) and inspiring subsequent research (Arnab et al., 2021; Neimark et al., 2021) in video understanding. Building on the ViT model, the Video Vision Transformer (ViViT) (Arnab et al., 2021) also represents videos as sequences of patches but differs from frame-based methods in that it employs a temporal tokenization approach. ViViT segments videos into fixed-length clips, representing each as a serie of temporal patches, which are linearly projected and processed through the transformer encoder, offering a refined strategy for analyzing video content. TimeSformer (Bertasius et al., 2021) also adapts the standard Transformer architecture to video by enabling spatio-temporal feature learning directly from a sequence of frame-level patches.

Applying standard self-attention mechanisms to video data, which involves comparing image patches from all spatial locations and frames, can lead to redundant spatial information at the expense of temporal dynamics. MotionFormer (Patrick et al., 2021) introduced a trajectory attention block specifically designed to enhance video transformers. This innovation

focuses on aggregating information along motion trajectories, effectively overcoming the computational and memory constraints tied to processing large inputs, thereby enhancing efficiency, especially with high-resolution content. (Wu et al., 2022a) introduced a memory-centric methodology for enhancing long-term video comprehension. This approach leverages the "keys" and "values" within a transformer's architecture as a form of memory, allowing queries to interact with an expanded dataset of keys and values sourced from both current and past sequences. By enabling each layer to reach further back into past data, the model achieves a significantly broader receptive field, using only half the parameters compared to the model proposed by (Patrick et al., 2021).

In an effort to refine multi-modal integration techniques, recent studies aim to develop models that perform consistently well across various modalities, rather than being excessively fine-tuned for individual ones. (Girdhar et al., 2022) unveiled a transformer-based framework that capitalizes on the adaptable nature of transformers. This model is simultaneously trained on classification tasks from disparate modalities, including 2D images, 3D images, and videos. (Yan et al., 2022) introduced a multi-view transformer tailored for multi-modal input, generating multiple "views" or representations by tokenizing spectrograms, optical flow, and RGB content using tubelets of varying dimensions. These tokens are processed through distinct encoders, integrated via a fusion module, and ultimately consolidated by a global encoder.

Another recent trend is to leverage over Large Language Models (LLMs), to obtain stronger representations. (Kazakos et al., 2021b) proposed a transformer-based multi-modal model that ingests video and audio as input modalities, with an explicit language model providing action sequence context to enhance the predictions.

Datasets

The availability of large-scale datasets has been instrumental in propelling the field of egocentric vision research forward. Given the relatively recent adoption of wearable cameras, there is a scarcity of egocentric video data available online. This section provides an overview of datasets employed for action recognition tasks, detailing their characteristics. It is important to note that the majority of these datasets are designed for general purposes, making them applicable to a wide range of tasks beyond action recognition. In Table 2.1 we conduct a comparative analysis of the most prominent publicly accessible egocentric datasets, examining their domains, size, and available modalities.

Dataset	Settings	Signals	Hours	Sequences	AVG. video duration	Participants
ADL (Pirsiavash and Ramanan, 2012)	Daily activities	RGB	10.0	20	30.00 min	20
UTE (Lee et al., 2012)	Daily Activities	RGB	37.0	10	222.00 min	4
EGTEA Gaze+[†] (Li et al., 2018e)	Kitchen	RGB, gaze	27.9	86	19.53 min	32
EPIC-KITCHENS-100 (Damen et al., 2022)	Kitchens	RGB, audio	100.0	700	8.57 min	37
MECCANO (Ragusa et al., 2023b)	Industrial	RGB, depth, gaze	6.9	20	20.79 min	20
Assembly101 (Sener et al., 2022)	Industrial	RGB, multi-view	167.0	1425	7.10 min	53
Ego4D[⋆] (Grauman et al., 2022)	Multi Domain	RGB, Audio, 3D, gaze, IMU, multi	3670.0	9650	24.11 min	931

Table 2.1 **General Egocentric Dataset - Collection Characteristics.** [†]: For EGTEA, Audio was collected but not made public. [⋆]: For Ego4D, apart from RGB, the other modalities are present for subsets of the data.

The **Activity of Daily Living (ADL)** dataset (Pirsiavash and Ramanan, 2012) emerged as one of the initial egocentric datasets. It includes one million frames recorded within home environments. Participants were given broad instructions to engage in everyday activities like watching TV or doing laundry, making the dataset minimally scripted. It includes annotations for object trajectories, hand positions, and interaction events. ADL has been utilized in a variety of research tasks, including action (Vondrick et al., 2016) and region anticipation (Furnari et al., 2017), action recognition (Pirsiavash and Ramanan, 2012), and video summarization (Lu and Grauman, 2013). Similarly, the **UTE** dataset (Lee et al., 2012) features video recordings from 4 participants engaged in diverse activities such as eating, shopping, attending lectures, driving, and cooking. A distinct characteristic of UTE, in comparison to ADL, is its video length; the average duration of a UTE video is 3.7 hours (222 minutes), significantly longer than the 30-minute average of an ADL video.

Differing in domains and captured signals, the **GTEA Gaze** dataset (Fathi et al., 2012b) and its extension **EGTEA Gaze+** (Li et al., 2018e) both center around recipe preparation within a single kitchen environment. The original GTEA Gaze dataset, introduced by (Fathi et al., 2012b), emphasizes action recognition and gaze prediction. It involves the use of eye-tracking glasses equipped with an inward-facing infrared gaze sensing camera to track the 2D location of subjects' eye gaze during meal preparation activities. The dataset comprises 17 sequences performed by 14 subjects following pre-specified meal recipes, annotated with 25 frequently occurring actions such as "take", "pour", and "spread" along with their respective starting and ending frames. This dataset was subsequently extended as EGTEA Gaze+ by (Li et al., 2018e), incorporating 28 hours of cooking activities. EGTEA Gaze+ includes video footage, gaze tracking data, action annotations for 106 actions, and pixel-level hand masks. It has been leveraged to address various tasks, including anticipation (Furnari and Farinella, 2019; Girdhar and Grauman, 2021; Zhong et al., 2023), action recognition (Fathi et al., 2012b; Kazakos et al., 2021b), procedural learning (Bansal et al., 2022), and future hand mask prediction (Jia et al., 2022).

Fig. 2.5 **EPIC-KITCHENS dataset.** Examples from the EPIC-KITCHENS dataset along with the corresponding actions.

While existing egocentric datasets offer valuable insights into various aspects of vision, their limited scale and focus on specific environments or individuals pose challenges when training deep learning models. In response, the **EPIC-KITCHENS** dataset (Damen et al., 2018) emerged in 2018 as a significantly larger egocentric video dataset, subsequently extended to the latest version, EPIC-KITCHENS-100 (Damen et al., 2022).

Comprising 100 hours of unscripted video recordings from 37 participants across 4 countries within their own kitchens, EPIC-KITCHENS stands out for its unique participant instructions. Participants are asked to perform recordings upon entering the kitchen and cease upon leaving, allowing for an unscripted exploration of their environments and the pursuit of personal goals. The dataset encompasses 90,000 action segments, 20,000 unique narrations, 97 verb classes, and 300 noun classes.

Recently, EPIC-KITCHENS has been further augmented with three additional annotations. Firstly, EPIC-KITCHENS Video Object Segmentations and Relations (VISOR) (Darkhalil et al., 2022) offers pixel-level annotations focusing on hands, objects, and hand-object interaction labels. VISOR includes 272,000 manual semantic masks of 257 object classes, 9.9 million interpolated dense masks, and 67,000 hand-object relations. Secondly, EPIC-SOUNDS (Huh et al., 2023) annotates temporally distinguishable audio segments from the video's audio stream, encompassing 78.4 thousand categorized segments of audible events and actions across 44 classes, along with 39.2 thousand uncategorized segments. Lastly, EPIC-Fields (Tschernezki et al., 2024) successfully registers and provides camera poses for 99 out of the 100 hours of EPIC-KITCHENS data.

Since its introduction, EPIC-KITCHENS has emerged as the default dataset for a wide range of egocentric vision tasks, including action recognition (Girdhar et al., 2022; Kazakos et al., 2019b; Xiong et al., 2022; Yan et al., 2022), privacy concerns (Thapar et al., 2020), and anticipation (Furnari and Farinella, 2019; Gu et al., 2021; Jia et al., 2022; Liu et al., 2020; Pasca et al., 2023; Roy and Fernando, 2022; Zhong et al., 2023).

Additionally, new research avenues have been opened up by EPIC-KITCHENS, particularly in the realm of domain adaptation, due to its diverse capture locations and temporal

Fig. 2.6 **Ego4D dataset.** Examples from the Ego4D dataset.

variability (Kim et al., 2021a; Munro and Damen, 2020a; Sahoo et al., 2021a), video retrieval (Lin et al., 2022; Zhao et al., 2023), manipulations (Shaw et al., 2023), as well as specialized topics such as object-level reasoning (Baradel et al., 2018) and learning words in other languages from visual representations (Surís et al., 2020).

A couple of datasets are tailored to industrial-like settings. **MECCANO**, introduced by (Ragusa et al., 2021, 2023b), is an egocentric procedural dataset capturing subjects assembling a toy motorbike model. This dataset includes synchronized gaze, depth, and RGB data, covering 20 object classes encompassing components, tools, and an instructions booklet. MECCANO has been leveraged for various tasks such as action recognition (Deng et al., 2023), active object detection (Fu et al., 2022), hand-object interactions (Tango et al., 2022), and procedural learning (Bansal et al., 2022).

Similarly, **Assembly101** (Sener et al., 2022) is a procedural activity dataset featuring 4,321 videos of individuals assembling and disassembling 101 "take-apart" toy vehicles. This dataset showcases diverse variations in action sequences, including mistakes and corrections. It contains over 100K coarse and 1M fine-grained action segments, along with 18M 3D hand poses. Assembly101 has been applied in action recognition (Wen et al., 2023), anticipation (Zatsarynna and Gall, 2023), and hand pose estimation (Ohkawa et al., 2023; Zheng et al., 2023).

The most remarkable and extensive dataset to date is **Ego4D** (Grauman et al., 2022). It comprises 3,670 hours of daily-life activity videos covering hundreds of unscripted scenarios, including household, outdoor, workplace, and leisure activities. These videos were captured by 931 unique camera wearers from 74 locations across 9 countries. Some examples from Ego4D are shown in Figure 2.6. The dataset primarily consists of videos, with additional subsets containing audio, eye gaze, and 3D meshes of the environment. Ego4D was released with a comprehensive set of benchmarks and annotations for train/val/test splits, focusing on past events (episodic memory queries), present activities (hand-object manipulation,

audio-visual conversation, social interactions), and future activities (activity and trajectory forecasting). Due to its massive scale and unconstrained nature, Ego4D has proven to be valuable for various tasks, including action recognition (Lange et al., 2023; Liu et al., 2022), action detection (Wang et al., 2023a), visual question answering (Bärmann and Waibel, 2022), active speaker detection (Wang et al., 2023b), natural language localization (Liu et al., 2023), natural language queries (Ramakrishnan et al., 2023), gaze estimation (Lai et al., 2022), persuasion modeling for conversational agents (Lai et al., 2023), audio-visual object localization (Huang et al., 2023), hand-object segmentation (Zhang et al., 2022a), and action anticipation (Mascaró et al., 2023; Pasca et al., 2023; Ragusa et al., 2023a). The diversity of Ego4D has led to the introduction of new tasks, such as modality binding (Girdhar et al., 2023), part-based segmentation (Ramanathan et al., 2023), long-term object tracking (Tang et al., 2024), relational queries (Yang et al., 2023), and action generalization across scenarios (Plizzari et al., 2023b). Moreover, its unprecedented scale has facilitated training robot models, leading to groundbreaking advancements in learning from demonstrations (Ma et al., 2022; Nair et al., 2023; Radosavovic et al., 2023). The potential of the Ego4D dataset is yet to be fully explored, and it continues to inspire research across multiple domains. Recently, the authors of Ego4D also introduced Ego-Exo4D (Grauman et al., 2023). This new dataset focuses on simultaneously captured egocentric and exocentric videos of skilled human activities such as sports, music, dance, and bike repair. It involves more than 800 participants from 13 cities around the world, performing these activities in 131 different natural scene contexts. This has resulted in long-form captures ranging from 1 to 42 minutes each, accumulating a total of 1,422 hours of video. The multi-modal nature of the dataset is unprecedented; it includes multi-channel audio, eye gaze, 3D point clouds, camera poses, IMU data, and multiple paired language descriptions. This also introduces a novel feature: "expert commentary" provided by coaches and teachers, specifically tailored to the domain of skilled activities.

In this thesis, we extensively utilize the EPIC-KITCHENS in Chapter 3, Chapter 5 and Chapter 6 and Ego4D dataset in Chapter 4. EPIC-KITCHENS has been collected in various kitchens, each corresponding to a different environment. These exhibit a significant domain shift, making them ideal for cross-domain analysis. In Ego4D, the shift is not limited to the environment but also encompasses the various scenarios presented and the geographical locations where activities are recorded. In Chapter 4, we introduce the Action Recognition Generalization Over scenarios and locations dataset (ARGO1M), which contains 1.1M video clips from the large-scale Ego4D dataset (Grauman et al., 2022), across 10 scenarios and 13 locations.

Limitations and future works

Despite the increasing interest in action recognition for egocentric videos, there are several areas that require attention from the computer vision community.

Firstly, there is a lack of approaches specifically tailored for egocentric vision. Many existing architectures are adapted from those designed for third-person videos and may not be optimized for the ego viewpoint or camera motion. Consequently, the ability to recognize fine-grained actions in egocentric videos lags behind that of third-person videos. Even with the utilization of transformer architectures and multiple modalities, state-of-the-art methods currently achieve only modest activity classification accuracy (e.g., 51.0% on EPIC-KITCHENS-100 by (Zhao et al., 2023)). It remains unclear whether this limitation stems from dataset size, label ambiguity, or the need for novel architectures.

Secondly, although the integration of gaze has shown promise for egocentric action recognition, subsequent datasets do not adequately capture the rich, albeit costly, egocentric gaze data. While a few sequences in Ego4D (Grauman et al., 2022) include gaze information, they are not specifically labeled with fine-grained actions. Gaze data can provide valuable insights into attentional focus and action anticipation. However, the lack of large-scale egocentric action recognition datasets with gaze information hinders further exploration in this area.

Lastly, the heavy reliance on labeled datasets for training poses limitations on model capabilities. Acquiring labeled data is not only costly but also subject to decisions regarding the choice of action classes and granularity. Transitioning from a closed subset to open labels remains a challenge in egocentric action recognition, as in many other machine learning tasks. With the emergence of Large Language Models (LLMs), the future of action recognition may lie in leveraging their capabilities, although metrics to assess success and monitor progress in this direction are currently missing.

2.1.3 3D Scene Understanding

The goal of 3D scene understanding is to teach an AI agent to interpret the surrounding environment and explore possible interactions with it. This involves understanding the environment itself, interactions that the user can perform within it, as well as objects in the scene and their locations. This field has attracted attention over the last few years, leading to the introduction of several new tasks and datasets. In this section, we first review current tasks

and methodologies for 3D scene understanding. We then introduce the datasets employed for this purpose. Finally, we present limitations of current approaches and future works.

Tasks and methodologies

The first work to make use of 3D information in egocentric videos is that of (Damen et al., 2014). Given a mapped environment, they used gaze estimation to cluster interaction regions into task-relevant objects in 3D and their modes of interaction. To examine interactions centred around humans within their surroundings, (Bertasius et al., 2015) suggested the use of egocentric stereo cameras. This technique sets up an egocentric object prior within an RGBD frame from a first-person perspective, which can be applied to detect 3D saliency. The work by (Rhinehart and Kitani, 2016) focused on learning and predicting of "action maps", which encode the user's capacity to carry out activities at different locations. This method maps actions to distinct areas within a scene, thereby facilitating the comprehension and anticipation of human activities within a particular environment. (Li et al., 2022b) focused on predicting the intended destination of a person's object manipulation action within a 3D workspace. Although this is a specific instance of trajectory forecasting, traditional methods are impractical in manipulation scenarios where the hands may not be visible in the camera's field of view. Consequently, focusing on the prediction of the 3D target location offers a clearer insight into potential interactions with objects, which is beneficial for applications like robotic planning and control. Recently, (Grauman et al., 2022) introduced the task of Visual Queries with 3D Localization (VQ3D), which aims to retrieve the relative 3D position of a queried object in relation to the current query frame. In this setting, different methods have been proposed. (Xu et al., 2023) introduced a transformer-based module that enhances the context of an object-proposal set by incorporating query information. Mai et al. (Mai et al., 2022) developed a framework that effectively combines 3D multi-view geometry with 2D object detection from egocentric videos, resulting in more accurate camera pose estimations and significantly better VQ3D outcomes. This method operates in three key stages: initially, a sparse 3D reconstruction is carried out using Structure from Motion (SfM) to extract 3D poses and generate a sparse 3D model. Subsequently, an egocentric video alongside a visual crop of the queried object is input into a model that detects relevant frames and their associated 2D bounding boxes. Finally, missing 3D poses in the identified frames are aligned with the sparse 3D model, and the 3D centroid movements of the object are calculated. The work of (Majumder et al., 2023) introduces another challenge: constructing a map of an unfamiliar 3D environment using the shared information found within the egocentric audio-visual observations of participants engaged in a natural conversation. More recently, (Nagarajan

and Grauman, 2020) proposed a reinforcement learning approach empowering an embodied agent to independently identify the affordance landscape within unfamiliar, unmapped 3D spaces, thereby facilitating the exploration of interactions. (Do et al., 2022) focused on forecasting the depths and surface normals of the surrounding environment based on a single-view egocentric image. They tackled the obstacles posed by wearable devices, like inclined images and dynamic objects in the foreground, by introducing an image stabilization technique. This method adjusts tilted images to a standard orientation, enhancing the learning process. (Nagarajan et al., 2024) proposed to learn environment-aware video representations that represent the surrounding physical space, aiming at facilitating the prediction of local environment states at different time-steps. They defined the local environment state in an egocentric video as the objects and their approximate distances in front, behind, to the left, and to the right of the camera-wearer. These states serve as training data for a transformer-based video encoder model, which gathers visual information across the entire video and constructs an environment memory. This memory enables the prediction of the local state at any designated point within the video. (Liu et al., 2022) introduced the challenge of simultaneously recognizing and locating the actions of a user within a pre-mapped 3D environment, using egocentric video footage. They designed an innovative deep probabilistic framework that employs a Hierarchical Volumetric Representation (HVR) of the 3D space, alongside the egocentric video, to infer the action's 3D location and recognise the action by leveraging contextual indicators. Finally, (Qian and Fouhey, 2023) tackled the challenge of predicting the 3D location, physical characteristics, and affordances of objects from single images. By processing a collection of query points, their method outputs predictions on potential 3D interactions, detailing aspects such as movability, location, rigidity, articulation, actions, and affordances. This is accomplished through a transformer-based model that enhances a detection backbone, providing a comprehensive understanding of objects and their potential interactions within a space.

Datasets

General-purpose egocentric datasets like **Ego4D** (Grauman et al., 2022) or **Ego-Exo4D** (Grauman et al., 2023), discussed in Section 2.1.2, serve as valuable resources for scene understanding tasks. However, there are also task-specific datasets designed to address specific challenges of 3D scene understanding in egocentric vision.

The **Egocentric Depth on everyday INdoor Activities (EDINA)** dataset introduced by (Do et al., 2022) aims to advance the understanding of dynamic egocentric scenes. It

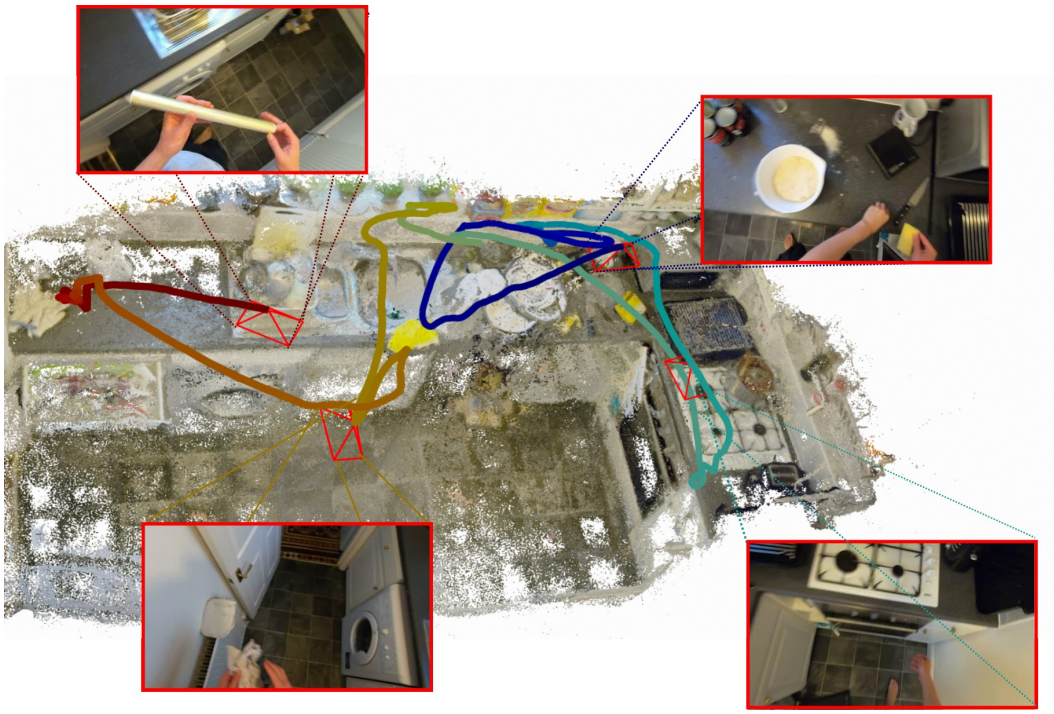


Fig. 2.7 **EPIC-Fields dataset.** The EPIC-Fields dataset (Tschernezki et al., 2024) provides 3D point clouds of the environments from EPIC-Kitchens recordings, along with the corresponding camera pose for each video frame. *Image from https://epic-kitchens.github.io/epic-fields*

comprises over 500K synchronised RGBD frames and gravity directions, covering a wide range of daily activities across 16 hours of RGBD recordings.

EgoPAT3D (Li et al., 2022b) provides a multi-modal dataset with over a million frames of RGB-D and IMU data. It is specifically designed for predicting the 3D target locations of object manipulation actions. This dataset includes 150 recordings, 15 household scene point clouds, 15,000 hand-object interactions, and 600 minutes of RGB-D/IMU footage, totaling 0.9 million hand-object action frames and 1 million RGB-D frames.

(Qian and Fouhey, 2023) introduced the **3D Object Interaction Dataset (3DOI)**, which incorporates internet videos, egocentric footage, and indoor images. It comprises 2K egocentric images selected from EPIC-KITCHENS (Damen et al., 2022) and annotated with 3D ground truths such as depth, surface normals, and interactable points on objects. The annotations include information about object movability, location, rigidity, articulation, potential actions, and affordances.

The **Aria Digital Twin** dataset (Pan et al., 2023), captured using Aria glasses, contains 200 sequences of real-world activities conducted by Aria wearers in two indoor scenes. It includes various data modalities such as raw camera streams, IMU streams, sensor calibra-

tion, ground truth data on device and object poses, eye gaze vectors, human poses, image segmentations, depth maps, and synthetic renderings.

(Ravi et al., 2023) proposed **ODIN (the OmniDirectional INdoor dataset)**, a large-scale dataset comprising over 300K omnidirectional images capturing diverse activities of daily living. It includes scans of the recording environments from a 3D scanner and camera-frame 3D human pose estimates, facilitating scene understanding tasks.

Recently, (Tschernezki et al., 2024) released **EPIC-Fields**, an extension of EPIC-KITCHENS with 3D camera poses. Covering 99 hours of recordings in 45 kitchens, EPIC-Fields reconstructs 96% of videos from EPIC-KITCHENS, offering opportunities to integrate 3D geometry into egocentric video understanding.

In this thesis, we leverage EPIC-Fields to enhance egocentric video models with 3D information (Chapter 6).

Limitations and future works

Egocentric video footage, by directly linking the actions of the camera wearer to their immediate 3D spatial surroundings, offers a distinct perspective in the field of egocentric vision. The integration of 3D scene models into this domain has been propelled forward by recent advances in 3D scanning technologies and the proliferation of head-mounted displays, facilitating the creation of rich datasets aimed at addressing 3D-centric research questions. However, challenges such as motion blur and unconventional captured angles, inherent to the nature of egocentric videos, pose obstacles to the 3D reconstruction process, especially for scenes with dynamic elements. This results in a notable gap in thoroughly comprehending 3D dynamics of movements and interactions within these environments. An intriguing direction for future investigation is the synthesis of egocentric (first-person) and exocentric (third-person) perspectives. Merging these viewpoints could unlock a more comprehensive understanding of intricate environments and human behaviors.

2.2 Learning across Domains

Despite the advancements in video analysis tasks, a common assumption in many existing methods is that training and test data share the same distribution. However, this assumption often fails in real-world settings, where the distribution of publicly available training data and real-world data frequently differs, leading to a *domain shift* between the training (source) and

testing (target) domains. This discrepancy results in the diminished effectiveness of video models in the target domain, despite the advanced capabilities of deep neural networks. To address the drop in model performance due to domain shifts, a variety of domain adaptation techniques have been developed. Unsupervised Domain Adaptation (UDA) focuses on adapting models from a labeled source domain to an unlabeled target domain by mitigating the impact of domain shifts without incurring high annotation expenses. Domain Generalization (DG), on the other hand, aims to build models that can generalize well across any unseen domain without requiring any access to data from these domains during the training phase. This section explores the methods developed for learning across domains, with a specific focus on UDA (Section 2.2.2) and DG (Section 2.2.3).

2.2.1 Problem Formulation

Unsupervised Domain Adapation. In UDA, we are given a collection of M_S source domains $\{D_S^1, D_S^2, \ldots, D_S^{M_S}\}$ and a collection of M_T target domains $\{D_T^1, D_T^2, \ldots, D_T^{M_T}\}$. Each source domain contains N_S^k videos and their corresponding labels $D_S^k = \{(V_S^{k,i}, y_S^{k,i})\}_{i=1}^{N_S^k}$, characterized by the underlying probability distribution p_S^k associated with the label space $\mathcal{Y}_S^k$ that contains $|C_S^k|$ classes. Each target domain contains N_T^j unlabeled videos $D_T^j = \{V_T^{j,i}\}_{i=1}^{N_T^j}$, characterized by the underlying probability distribution p_T^j associated with the label space $\mathcal{Y}_T^j$ that contains $|C_T^j|$ classes, such as $|C_S^k|=|C_T^j|$. The goal of UDA is to design a target model which is capable of learning transferable features from the labeled source domains and minimize the empirical target risk ϵ_T across all target domains performed on certain tasks.

Domain generalization. Domain Generalization (DG) aims to learn a model from one or multiple source domains that can generalize well to unseen target domains without accessing any data from them. In other words, the goal of DG is to learn a model using only the data from the source domains D_S^k so that the model performs well on any target domain D_T^k. The model should be capable of learning domain-invariant features from the labeled source domains in order to minimize the empirical target risk ϵ_T on the target domain. DG is primarily explored in two frameworks: multi-source DG and single-source DG. Multi-source DG operates on the assumption that multiple distinct domains are available for training (i.e., $N_s^k>1$), with the aim of leveraging this variety to learn domain-invariant representations. This approach is crucial, especially since models lack direct access to target domain data, which poses a challenge for generalization. By utilizing multiple domains, models can find stable patterns that are more likely to generalize effectively to new, unseen domains. On the other

hand, single-source DG assumes a homogeneous training set derived from a single domain (i.e., N_s^k=1). This problem is closely related to the topic of Out-of-Distribution (OOD) robustness (Hendrycks and Dietterich, 2018), which investigates model robustness under image corruptions. Despite this distinction, most methods do not explicitly align themselves with either single-source or multi-source DG, opting instead for a broader approach to generalization, and are tested across datasets that include both single- and multi-source environments.

2.2.2 Unsupervised Domain Adaptation

As for Unsupervised Domain Adaptation (UDA) methods, which leverage unlabeled target data during training, they can be broadly categorized into two main categories.

Discrepancy-based methods aim to minimize a distance metric between the source and target distributions (Long et al., 2015; Saito et al., 2018; Xu et al., 2019a). Those are the Maximum Mean Discrepancy (MMD) (Long et al., 2015), Correlation Alignment (CORAL) (Sun and Saenko, 2016), Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), and Contrastive Domain Discrepancy (CDD) (Kang et al., 2019). *Adversarial-based methods* use adversarial training to align source and target distributions (Deng et al., 2019; Tang and Jia, 2020). The basic idea of these methods is to incorporate a domain classifier trained to predict the domain of the input data, i.e., whether they come from the source or the target domain. A key component of these methods is the use of a Gradient Reversal Layer (GRL), which reverses the gradients flowing from a domain classifier to the feature extractor (Ganin and Lempitsky, 2015a) so that the feature extractor can learn domain-invariant features (see Figure 2.8). Other works exploit batch normalization layers to normalize source and target statistics (Chang et al., 2019; Li et al., 2017c, 2018d). Authors of AdaBN (Li et al., 2018d) show that domain-related knowledge is represented by the statistics of the Batch Normalization (BN) (Ioffe and Szegedy, 2015) layers. Therefore, they achieve transfer of the trained model to a new domain by modulating the statistics in the BN layer. An alternative research direction incorporates self-supervised learning as an auxiliary task to improve feature learning, as in (Bucci et al., 2021).

While the approaches mentioned above have primarily been utilized for standard image classification tasks, there has also been a substantial amount of research focused on Unsupervised Domain Adaptation (UDA) for video-related tasks, such as action detection (Agarwal et al., 2020), segmentation (Chen et al., 2020), and classification (Chen et al., 2019; Choi et al., 2020b; Jamal et al., 2018; Munro and Damen, 2020a; Pan et al., 2020; Song et al.,

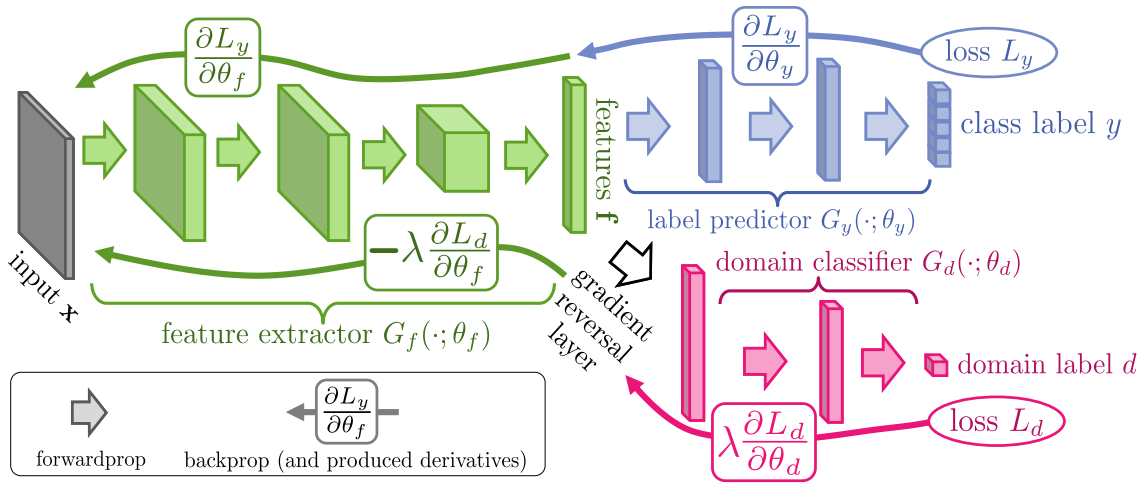


Fig. 2.8 **DANN architecture.** The architecture proposed in (Ganin and Lempitsky, 2015a) includes a deep feature extractor (green) and a label predictor (blue), which together form a standard feed-forward architecture. Unsupervised domain adaptation is achieved by adding a domain classifier (red) connected to the feature extractor via a gradient reversal layer that multiplies the gradient by a certain negative constant during the backpropagation-based training. Training is performed by minimizing the label prediction loss (for source examples) and the domain classification loss (for both source and target samples). Gradient reversal ensures that the feature distributions over the two domains are made similar (as indistinguishable as possible for the domain classifier), thus resulting in the domain-invariant features. *Image from (Ganin and Lempitsky, 2015a).*

2021b). In video classification, numerous Video Unsupervised Domain Adaptation (VUDA) methods have been introduced to align the temporal dynamics of the feature space. An overview of existing VUDA methods is provided in the following. We also summarize them in Table 2.2. We refer to the survey in (Xu et al., 2022) for a more detailed overview of VUDA methods.

Adversarial-based VUDA methods. Methods under this category leverage domain discriminators to identify whether videos come from the source or the target domain. Through adversarial objectives, the discrepancy between source and target domains is minimized impliciteply. Deep Adversarial Action Adaptation (DAAA) (Jamal et al., 2018) extends the original image-based DANN (Ganin and Lempitsky, 2015a) to videos, adapting both spatial and temporal features. Temporal Attentive Adversarial Adaptation Network (TA^{3}N) (Chen et al., 2019) uses a multi-level adversarial framework with temporal relation and attention mechanisms to achieve transfer from the source to the target domain. It leverages a Temporal Relation Network (TRN) (Zhou et al., 2018) to obtain more explicit temporal features, and align videos with both spatial and temporal features. Temporal Co-attention Network (TCoN) (Pan et al., 2020) adapts target video features to the source ones by constructing

Categories	Brief Descriptions	Methods
Adversarial-based	Domain discriminators to encourage domain confusion through adversarial objectives across video domain.	DAAA (Jamal et al., 2018), TA^{3}N (Chen et al., 2019), TCoN (Pan et al., 2020), MM-SADA (Munro and Damen, 2020a), CIA (Yang et al., 2022a)
Discrepancy-based	Discrepancy between domains are explicitly computed, align domains by applying metric learning approaches.	AMLS (Jamal et al., 2018)
Semantic-based	Domain-invariant features are obtained by exploiting the shared semantics across domains.	STCDA (Song et al., 2021b), CMCo (Sahoo et al., 2021b), CoMix (Sahoo et al., 2021b), CO^{2}A (da Costa et al., 2022), A^{3}R (Zhang et al., 2022d)
Reconstruction-based	Domain-invariant features from encoder-decoder networks with data-reconstruction objectives.	TranSVAE (Wei et al., 2022)
Composite	Exploit a composite of approaches to capitalise on the strength of each approach.	NEC-Drone (Choi et al., 2020a), SAVA (Choi et al., 2020c)

Table 2.2 Different categories of methods for closed-set VUDA. Methods are listed in chronological order.

target-aligned source features via transforming the original source video features through a cross-domain co-attention matrix. Besides spatial and temporal features which are generally obtained from the RGB modality, videos also contain information about other modalities, such as optical flow and audio modalities. The multi-modal nature of videos can improve VUDA methods as domain shift affects each modality differently. Methods have therefore been proposed to align source and target videos leveraging on multi-modal information. Among these, MM-SADA (Munro and Damen, 2020a) leverages the RGB and optical flow modalities, and applies adversarial alignment to each modality separately. MM-SADA further adopts self-supervised learning across different modalities to learn the temporal correspondence between them. Cross-modal Interactive Alignment (CIA) (Yang et al., 2022a) aligns video features with RGB, optical flow, and audio modalities. CIA further observes that cross-modal alignment could conflict with cross-domain alignment in VUDA, therefore it enhances the transferability of each modality by cross-modal interaction through a Mutual Complementarity (MC) module. The different modalities are therefore refined by absorbing the transferable knowledge from other modalities before they are aligned across source and target domains.

Discrepancy-based VUDA methods. Methods under this category tackle VUDA by computing and minimizing the domain discrepancy between source and target domains explicitly.

An early method is AMLS (Jamal et al., 2018) where the target videos are modeled as a sequence of points on the Grassmann manifold (Turaga et al., 2008) with each point corresponding to a collection of clips aligned temporally, and the source videos are modeled as a single point on the manifold. VUDA is tackled by minimizing the Frobenius norm (Huckle and Kallischko, 2007) between the source point and the series of target points on the Grassmann manifold.

Semantic-based VUDA methods. These methods rely on the shared semantics across the source and target domains to obtain domain-invariant features. Under this category is Spatio-Temporal Contrastive Domain Adaptation (STCDA) (Song et al., 2021b), which extracts video representations from both RGB and optical flow modalities by employing a contrastive loss at both clip and video levels, ensuring that frames and clips are aligned both spatially and temporally. STCDA further addresses the domain shift between source and target videos through a Video-based Contrastive Alignment (VCA) loss, which reduces the distance between intra-class features of source and target while increasing the distance between inter-class features. The labels for target videos are assigned through a pseudo-labeling process, which involves clustering based on the features of the labeled source videos. Contrastive learning has also been applied in CMCo (Sahoo et al., 2021b) to extract video features with modality correspondence across RGB and optical flow modalities. Similarly, CO^{2}A (da Costa et al., 2022) trains video feature extractors with the goal of achieving feature clustering through contrastive learning, applied at both clip and video levels. CO^{2}A further integrates supervised contrastive learning (Khosla et al., 2020) into the learning process for source video features. CoMix (Sahoo et al., 2021b) uses contrastive learning to enforce temporal speed invariance in videos by encouraging features extracted from the same video yet sampled with different temporal speeds to be similar. The authors of A^{3}R (Zhang et al., 2022d) note that the sounds produced by actions can serve as natural cues that are invariant across domains. They introduce a mechanism for learning about activities that are not present, utilizing audio-based predictions to identify actions that are inaudible in a video. Concurrently, they encourage visual predictions to assign low probabilities to these "pseudo-absent" actions. A^{3}R also implements an audio-balanced learning strategy, utilizing audio from the source domain to cluster samples.

Reconstruction-based VUDA methods. These methods address VUDA by extracting domain-invariant features using an encoder-decoder network, which is trained with data-reconstruction objectives. Several image-based domain adaptation studies have been using the reconstruction-based approach (Deng et al., 2021; Ghifary et al., 2016; Yang et al.,

2020), benefiting from its resilience to noise. However, there have been limited efforts in adapting reconstruction-based methods for videos, due to the challenges associated with video reconstruction. TranSVAE (Wei et al., 2022) is a recent attempt in leveraging data-reconstruction objectives for VUDA. It aims to disentangle domain-specific information from domain-invariant information during adaptation by generating cross-domain videos from two sets of latent factors, one encoding the static information and another encoding the dynamic information. This is done through a Variational AutoEncoder (VAE) (Kingma et al., 2019).

Composite of approaches. To exploit the strength of each approach for a more effective VUDA, various VUDA methods exploit a composite of approaches. For example, (Choi et al., 2020a) proposed to combine an adversarial-based approach with a semantic-based approach in a challenging setting where the label sets from source and target domains are different. SAVA (Choi et al., 2020c) aligns source and target video domains adversarially while encouraging temporal association in videos through an auxiliary clip order prediction task.

2.2.3 Domain Generalization

Previous approaches for Domain Generalization (DG) are mostly designed around image data (Bucci et al., 2021; Carlucci et al., 2019; Dou et al., 2019; Li et al., 2018b,c; Volpi et al., 2018). Most existing image-based DG approaches fall into the category of **domain alignment** (Ganin et al., 2016; Li et al., 2018b; Sun and Saenko, 2016; Yang et al., 2022b), where the core idea is to minimize the differences among source domains to learn domain-invariant representations. The rationale is straightforward: features that are invariant to the source domain shift should also be robust against any unseen target domain shift. This can be achieved by minimizing distances such as Maximum Mean Descrepancy (MMD) (Gretton et al., 2012; Li et al., 2018b). Utilizing an autoencoder architecture, (Li et al., 2018b) minimized the MMD distance between source domain distributions for hidden-layer features. As a commonly used distribution divergence measure, the Kullback-Leibler (KL) divergence has also been employed for domain alignment (Li et al., 2020; Wang et al., 2021c). Different from explicit distance measures like the MMD and KL divergence, adversarial learning (Jia et al., 2020; Li et al., 2018c; Matsuura and Harada, 2020; Rahman et al., 2020) formulates the distribution minimization problem through an adversarial discriminator (Ganin et al., 2016). In DG, adversarial learning is performed between source domains to learn source domain-agnostic features that are expected to work in novel domains (Li et al., 2018c; Rahman et al., 2020). Simply speaking, the learning objective is to make features confuse a domain

discriminator, which can be implemented as a multi-class domain discriminator (Matsuura and Harada, 2020), or a binary domain discriminator in a per-domain basis (Jia et al., 2020; Li et al., 2018c; Shao et al., 2019).

Data augmentation has been a common practice to regularize the training of machine learning models to avoid overfitting and improve generalization, which is particularly important for over-parameterized deep neural networks (Chen et al., 2022a,b; Nam et al., 2021; Volpi and Murino, 2019; Volpi et al., 2018; Wang et al., 2020c; Xu et al., 2020; Zhang et al., 2022b; Zhou et al., 2022). In doing so, the original data distribution is expanded, allowing the model to learn more generalizable features. Inspired by adversarial attacks (Goodfellow et al., 2014; Szegedy et al., 2013), several data augmentation methods use adversarial gradients obtained from the task classifier to perturb the input images (Qiao et al., 2020; Volpi et al., 2018). In (Mancini et al., 2020), Mixup (Zhang et al., 2018) is applied to mix instances of different domains in both pixel and feature space. MixStyle (Zhou et al., 2022, 2023) achieves style augmentation by mixing CNN feature statistics between instances of different domains.

Meta-learning aims to learn from episodes sampled from related tasks to benefit future learning (see (Hospedales et al., 2021) for a comprehensive survey on meta-learning). The motivation behind applying meta-learning to DG is to expose a model to domain shift during training with the hope that the model can better deal with domain shift in unseen domains (Balaji et al., 2018; Dou et al., 2019; Li et al., 2018a, 2019a,b) Existing meta-learning DG methods can only be applied to multi-source DG where domain labels are provided. The meta-learning paper most related to DG is MAML (Finn et al., 2017), which divides training data into meta-train and meta-test sets, and trains a model using the meta-train set in such a way to improve the performance on the meta-test set. In (Finn et al., 2017), MAML was used for parameter initialization, i.e., to learn an initialization state that is only a few gradient steps away from the solution to the target task.

Self-supervised learning is often referred to as learning with free labels generated from data itself (see (Jing and Tian, 2020) for a comprehensive survey on self-supervised learning). This can be achieved by teaching a model to predict the transformations applied to the image data, such as the shuffling order of patch-shuffled images (Bucci et al., 2021; Carlucci et al., 2019) or rotation degrees (Gidaris et al., 2018). An intuitive explanation is that solving pretext tasks allows a model to learn generic features regardless of the target task, and hence less over-fitting to domain-specific biases (Bucci et al., 2021).

Another recent trend is to learn **domain prompts** from visual (Shu et al., 2022; Zheng et al., 2022) or text information (Niu et al., 2022; Zhang et al., 2021), or utilize cross-modal

supervision (Min et al., 2022). For example, DoPrompt (Zheng et al., 2022) learns domain-specific prompts and trains a prompt adapter to generate a combination of these for each training image. The adapter is then used at test time to integrate knowledge from the source domains for each target image.

There are limited works on **video domain generalization**. VideoDG (Yao et al., 2021) highlights importance of striking a balance between generalization and discrimination, emphasizing the need for relationships between frames in the source domain to extend in a manner that facilitates generalization to potential target domains while preserving discriminative capabilities. To achieve this objective, an Adversarial Pyramid Network (APN) is employed, trained with adversarial data augmentation.

In this thesis, we introduce an approach for enhancing multi-modal video domain generalization by utilizing multi-modal features (Chapter 3). Recognizing that simple fusion of multi-modal data may not suffice for improving generalizability, we develop a unique cross-modal Relative Norm Alignment (RNA) loss. This loss function aligns the relative norms from multiple modalities and from various source domains, facilitating the generation of domain-invariant representations. We also propose a new method for domain generalization that represents each video as a weighted combination of other videos in the batch, potentially from different domains (Chapter 4). We name this method Cross-Instance Reconstruction (CIR), which is regulated by both a classification loss and a video-text association loss, paving the way for more robust and generalizable multi-modal learning.

2.3 Event-Based Cameras

This section introduces event-based cameras, a class of neuromorphic vision devices inspired by the efficient and asynchronous information processing observed in biological systems. Unlike traditional cameras that capture images at fixed intervals, often leading to redundant information, event-based cameras operate on a different principle. They mimic the human retina's behavior of responding only to changes in light intensity, thereby generating data in a sparse and energy-efficient manner. Each pixel in these cameras functions independently, detecting changes in brightness and triggering signals only when a significant brightness variation occurs. This approach not only reduces redundancy but also allows for real-time processing, capturing dynamic scenes with high temporal resolution. By adopting the biological retina's asynchronous signaling mechanism, event-based cameras offer a promising avenue towards developing vision systems that closely resemble the efficiency and precision of their human-like counterparts.

In Section 2.3.1, we delve into the foundational principles of neuromorphic vision devices, exploring the technological and biological inspirations behind their design and functionality. Section 2.3.2 discusses the methodologies for representing event-based data in a format that can be effectively processed by conventional deep learning neural networks. Finally, in Section 2.3.3, we introduce the primary datasets available for event-based vision research, providing a comprehensive overview of the resources that fuel advancements in this cutting-edge field.

2.3.1 Neuromorphic Vision Devices

Situated at the back of the ocular globe, the retina is an intricate network of neurons which is the foundation of humans' biological visual system. This multilayered structure contains specialized cells that transform light into neural signals. In the innermost membrane layer, pigment molecules capture incoming light, initiating the visual process. These molecules are sensitive to various light wavelengths, giving us the perception of a spectrum of colors. The interaction of light with these molecules induces a cascade of chemical reactions, altering the membrane potential of the photoreceptors known as rods and cones, which reside in the retina's inner layer. Bipolar cells act as intermediaries between these photoreceptors and the ganglion cells in the external synaptic layer. The latter generate action potentials that travel along the optic nerve. The structural design of the human retina is depicted in Figure 2.9.

The retina's ganglion cells encode visual information into patterns of action potentials, referred to as spike-trains. Contrary to common belief, these patterns do not directly represent the intensity of light or color. Rather, they are associated with the presence of motion and variations in brightness. Specifically, as changes in light intensity become more pronounced, the frequency of the ganglion cells' spikes also rises, and conversely, it diminishes in the absence of visual changes. These spikes are then transformed into continuous signals within the initial stages of the visual cortex, paving the way for advanced visual comprehension.

Inspired by the operational efficiency of natural neural systems, researchers have started to develop innovative architectures and algorithms. These designs emulate neurobiological processes, inherenting computational and communicative efficiencies of biological systems, and are referred to as *neuromorphic devices* or *event-based cameras*. Analogous to the neural layers in the biological retina that generate spikes in response to brightness changes, event-based devices produce an "event" in response to changes in light. These sensors consist of a pixel matrix that independently track the brightness level hitting on their photodiodes. Whenever the logarithmic intensity $L = \log(I)$ at pixel location (x_i, y_i) changes of a quantity

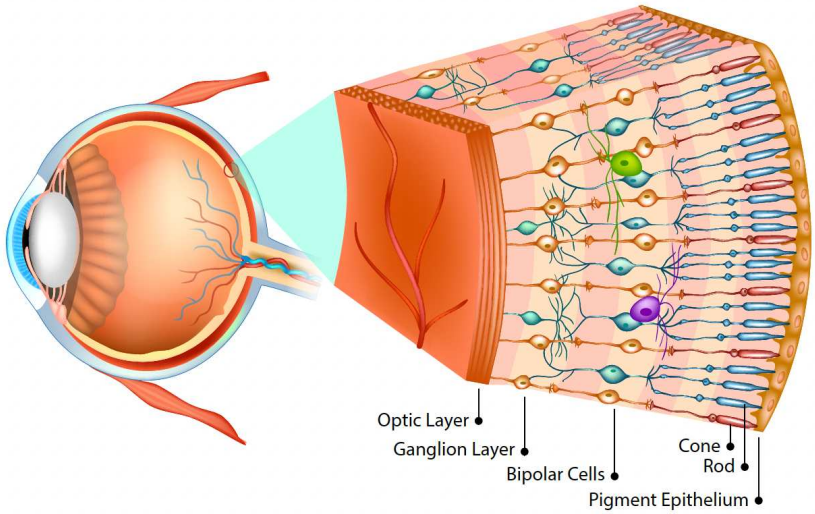


Fig. 2.9 **Cross section of the human eye's retina.** The light striking the retina travels through all the neural layers before reaching and activating the innermost rods and cones. These photoreceptors initiate communication back toward the ganglion cells and eventually through the optic nerve. *Image taken from https://hdl.handle.net/10589/187047*

above or below a predefined logarithmic threshold $C > 0$, an event

$$e_i = \{x_i, y_i, t_i, p_i\} \tag{2.1}$$

is triggered, capturing the pixel's position (x_i, y_i), the time t_i at which the change is detected, and a polarity bit $p_i \in \{-1, 1\}$ indicating whether the intensity decreased or increased. Two consecutive events e_i and e_j originating from the same pixel (x, y) adhere to the relation:

$$\Delta L(e_i, e_j) = L(x, y, t_i) - L(x, y, t_j) = p_i \cdot C, \quad \text{with} \quad \Delta t_{ij} = t_i - t_j > 0, \tag{2.2}$$

where Δt_{ij} represents the temporal interval between events, and L denotes the logarithmic brightness intensity. The output from these sensors is a sequence of asynchronous events $E = \{e_i | t_i > t_j \; \forall i > j\}$. The frequency of these events is intrinsically correlated to the dynamics within the scene; high rates of events are indicative of rapid movements, whereas fewer events are produced in more static scenarios. A comparison between the output of a traditional camera and that of an event-based camera is shown in Figure 2.11.

The *Dynamic Vision Sensor (DVS)* was first introduced by (Lichtsteiner et al., 2008) and then later improved by (Serrano-Gotarredona and Linares-Barranco, 2013), who increased the sensitivity of the pixels and reduced their size at the expense of higher power consumption. It operates by simulating the components of a biological retina, such as photoreceptors

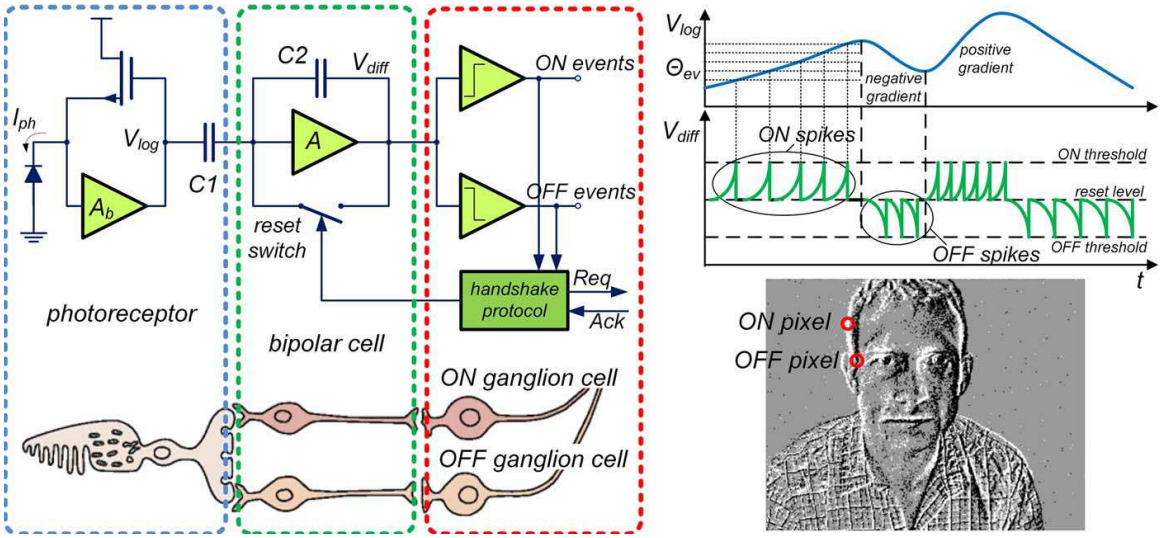


Fig. 2.10 Three-layer model of a human retina and corresponding Dynamic Vision Sensor (DVS) pixel circuitry are depicted on the left. The typical signal waveforms of the pixel circuit are illustrated in the top right panel. The upper trace shows a voltage waveform at the node V_{log}, which tracks the photocurrent through the photoreceptor. The bipolar cell circuit generates spike events (V_{diff}) of different polarities in response to both positive and negative changes in photocurrent. These spikes are then monitored by the ganglion cell circuit, which also transmits the spikes to subsequent processing stages. The magnitude of log-intensity change is encoded in the number of events, and the rate of change is indicated by the intervals between events. The bottom right image demonstrates the response of a DVS pixel array to a natural scene (a person moving within the sensor's field of view). Events, collected over tens of milliseconds, are displayed as an event map image, with ON events (increases in brightness) and OFF events (decreases in brightness) represented as white and black dots, respectively. *Image taken from* (Posch et al., 2014).

and bipolar and ganglion cells, through specialized hardware circuits. These circuits are responsible for the replication of the retina's processing capabilities. The sensor reacts to changes in brightness through a voltage signal $V_{\log}$, which is proportional to the logarithm of the incident light intensity. This signal is then amplified to a differential signal V_{diff}, which, when exceeding a certain threshold, generates an "event". An overview of the sensor is provided in Figure 2.10.

More advanced event-based camera designs propose combining brightness change detection with the direct measurement of pixels' intensity values, thus exploiting the benefits of both vision paradigms and more faithfully reproducing information available in the primary visual cortex. The *Asynchronous Time-Based Image Sensor (ATIS)* developed by (Posch et al., 2010) is the first device to provide this sort of combined visual information. This extends a traditional DVS pixel with an additional exposure measurement circuit, enabling

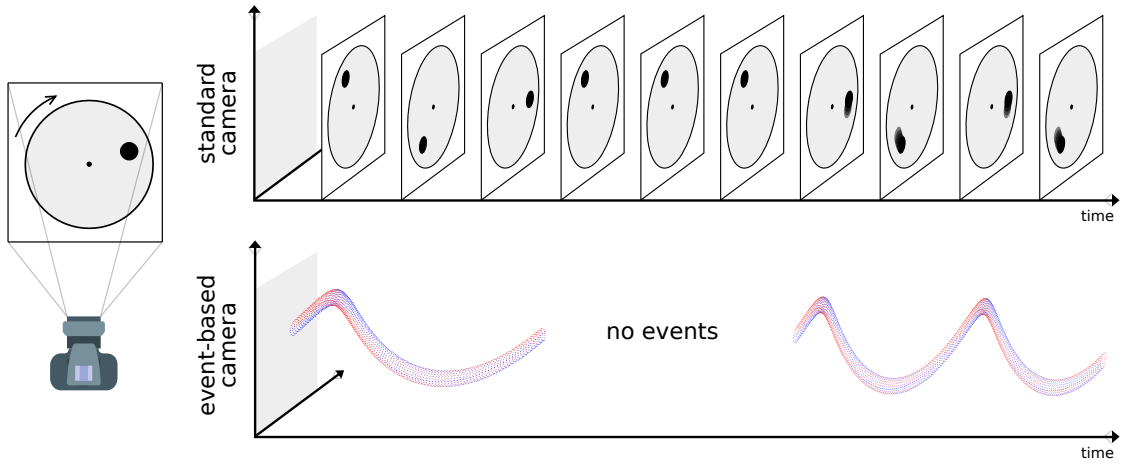


Fig. 2.11 **Standard camera vs event-based camera.** Comparison between the output of a standard camera and that of an event-based camera capturing a rotating disk. While a standard camera captures full-frame images at predefined intervals, an event camera only records changes in the scene. As a result, the background of the disk is not captured since its intensity does not change, thereby significantly reducing information redundancy. Similarly, no event is generated when the disk stops. Additionally, the high temporal resolution of an event camera eliminates motion blur effects, which would otherwise affect standard devices in scenarios involving high-speed motion. *Image taken from https://hdl.handle.net/10589/187047*

the recording of additional events encoding exposure measurements (EM events) analogous to that conveyed by grayscale images.

The integration of asynchronous DVS readouts with intensity measurements allows ATIS sensors to achieve high video compression rates and exceptional temporal resolution. However, the encoding time is inversely related to the intensity, potentially leading to artifacts with dark objects and disruptions by new events. Additionally, the pixel size imposes a fundamental limit on the resolution. A novel hybrid solution presented in the *Dynamic and Active pixel Vision Sensor (DAVIS)* offers a significant step forward. (Berner et al., 2013) developed DAVIS to provide both asynchronous and synchronous information, unlike the ATIS sensor. DAVIS uniquely fuses frame-based intensity readings with asynchronous events, enabling full-frame grayscale images and detecting brightness changes with high efficiency. It utilizes a single pixel for both operations, with traditional shutter mechanisms and event detection capabilities.

Event-based cameras provide several advantages over traditional cameras. Those are:

- *Low latency and temporal resolution:* DVS sensors are equipped with high-speed analog circuits capable of identifying changes in brightness with microsecond pre-

cision. As they operate asynchronously, they transmit data with minimal delay, in contrast to traditional cameras that require a global exposure time. This characteristic enables event-based cameras to record fast movements clearly, avoiding the motion blur typically associated with traditional frame-based cameras.

- *High Dynamic Range (HDR):* event-based cameras, unlike standard cameras that operate with a predetermined exposure time, do not adhere to a single exposure level. This flexibility allows them to function efficiently across a broad dynamic range of over 120 dB. As a result, they are capable of adjusting to different lighting environments, maintaining steady performance through their logarithmic reaction to changes in light intensity.

- *Low power consumption:* event-based cameras utilize power solely upon detecting changes, resulting in substantial energy efficiency. This means their power consumption can be remarkably low, often just around 100 mW for many models, which makes them especially ideal for use in wearable technology and mobile robots.

2.3.2 Deep Learning Approaches to Event Cameras

While event-based cameras offer several advantages, developing algorithms for effectively processing them poses some challenges. Events are asynchronous and spatially sparse, in contrast to the dense and rich information formats required by traditional vision algorithms. The two main approaches for tackling these challenges are: (i) *event-by-event computation*, leveraging the temporal dynamics of event data, and (ii) *grid-based representations*, which adapts event data for compatibility with conventional deep learning frameworks. The section offers an overview of current methods in both categories, highlighting how they address the unique properties of event data.

Event-by-event Computation

Event-driven processing techniques handle each event individually, updating the system's output progressively and asynchronously as events come in, thus ensuring minimum response times. Such algorithms typically maintain an evolving internal state, which is refreshed with each new event. *Spiking Neural Networks (SNNs)* (Maass, 1997b) are the leading approach of asynchronous, spike-based neural computation. These networks consist of neuron-like entities that process incoming spike events in an independent manner, firing when they have gathered sufficient relevant data. The membrane potential of these neurons forms

the SNNs' internal memory, which is dynamically updated with the arrival of new events. SNNs have been employed in various event-driven tasks, including edge detection (Meftah et al., 2010; Wu et al., 2007), object classification (Diehl et al., 2015; Lee et al., 2016), and gesture recognition (Botzheim et al., 2012). While they are traditionally trained using unsupervised biologically inspired learning rules (Hao et al., 2020; Rathi et al., 2018), they often exhibit enhanced performance with the incorporation of conventional gradient-based learning techniques. Several works use Artificial Neural Networks (ANNs) as intermediaries to learn synaptic weights to overcome SNNs' inherent non-differentiability (Diehl et al., 2015; Pérez-Carrasco et al., 2013; Rueckauer et al., 2017). *Filtering algorithms* constitute another principal category of event-driven computation methods. These algorithms are designed to work with partial and potentially noisy data, continuously updating a defined state with each new observation. This makes them a natural fit for asynchronous event-driven computation. Consequently, numerous event-driven filtering algorithms, both deterministic and probabilistic, have been developed, with applications ranging from Simultaneous Localization and Mapping (SLAM) (Gallego et al., 2017; Kim et al., 2008; Reinbacher et al., 2017) to noise filtering (Czech and Orchard, 2016; Khodamoradi and Kastner, 2018) and image ad video reconstruction (Munda et al., 2018; Scheerlinck et al., 2018), transforming the outputs of event cameras into more traditional visual formats. For event-based artificial neural networks, deterministic filters have been introduced to execute asynchronous convolution, facilitating feature extraction with high efficiency (Pérez-Carrasco et al., 2013; Scheerlinck et al., 2019). They exploit event cameras' sparse representation to conduct rapid computations on local areas triggered by events, avoiding the need to process entire images.

Grid-like Event Representations

Consider a sequence of asynchronous events defined by $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i)\}_{i=1}^{N}$ covering a time span ΔT. The procedure to derive a grid-based representation can be represented as a function Φ_R, which transforms $\mathcal{E}$ into a three-dimensional structure R_c within $\mathbb{R}^{H\times W\times F}$, where each pixel is characterized by F features. Over the past years, several methodologies for extracting these representations have been introduced. Typically, these grid-like representations are hand-crafted, which means the mapping function that converts the event stream to R_c is independent of the specific task. More recent developments (Cannici et al., 2020a; Deng et al., 2020a; Gehrig et al., 2019b) suggested the integration of neural network layers within Φ_R to be trained together with the entire network, aiming to extract representations that are task-specific. We provide an overview of the most popular representations in the

Representation	Dimensions	Description	Characteristics
Event frame	$H \times W$	Image of event polarities	Discards temporal and polarity information
Event count image	$2 \times H \times W$	Image of event counts	Discards time stamps
SAE	$2 \times H \times W$	Image of most recent time stamp	Discards earlier time stamps
Voxel grid	$B \times H \times W$	Voxel grid summing event polarities	Discards event polarity
HATS	$2 \times H \times W$	Histogram of average time surfaces	Discards temporal information
EST	$2 \times B \times H \times W$	Sample event point-set into a grid	Discards the least amount of information

Table 2.3 **Event-based representations.** Comparison of grid-based event representations used in prior work on event-based deep learning. H and W denote the image height and width dimensions, respectively, and B the number of temporal bins.

following, focusing on those that have been used in previous works as the input of deep neural networks. Those are summarized in Table 2.3.

Simple aggregation methods. Early deep neural network applications to event-based cameras have utilized elementary aggregation methods to process event data. Within such frameworks, a set of events $\mathcal{E}(x,y,p) = \{e_i \in \mathcal{E} \mid x_i = x, y_i = y, p_i = p\}$, categorized by polarity p and pixel location (x,y), is typically condensed into a single pixel matrix via basic aggregation techniques. The *event counts* model (Maqueda et al., 2018; Zhu et al., 2018) employs the cardinality $|\cdot|$ of $\mathcal{E}(x,y,p)$ to aggregate sequences of events, discarding any temporal information:

$$R_{\mathcal{E}}^{count}(x,y,p) = |\mathcal{E}(x,y,p)|.$$

Other representations use the event polarities and aggregate them into a two-dimensional *Event Frame* (Rebecq et al., 2017). In contrast, the *Surface of Active Events (SAE)* (Benosman et al., 2013; Zhu et al., 2018) retains only the most recent temporal information by recording the of the timestamp t_i of the last event received at each pixel:

$$R_{\mathcal{E}}^{SAE}(x,y,p) = \max_{i \in \mathcal{E}(x,y,p)} t_i,$$

where, for simplicity, we denote by $i \in \mathcal{E}(x,y,p)$ the indices of the events e_i in the sequence.

Voxel-grid based representations. A *voxel grid* representation (Zhu et al., 2019b) segments the event stream into a spatio-temporal grid $H \times W \times B$ by maintaining the original spatial resolution but dividing time into B consecutive bins. Two principal methodologies have been suggested (Wang et al., 2019a; Zhu et al., 2019b) to define the bins. The first method divides the time frame ΔT into B equally-sized sub-windows for which the representations R_b are derived, with each aggregating the subset $\mathcal{E}_b = \{e_i \in \mathcal{E} \mid t_i \in [(b-1)\frac{\Delta T}{B}, b\frac{\Delta T}{B}]\}$,

which corresponds to a singular $H \times W$ slice of the voxel grid. The alternative approach is analogous but constrains the quantity of events in each bin to a predetermined number N_e, thus partitioning the sequence into $\mathcal{E}_b = \{e_i \in \mathcal{E} \mid i \in [(b-1)N_e, bN_e]\}$ intervals. An example of a voxel grid representation is shown in Figure 2.12b. In this thesis, we use voxel-images from (Zhu et al., 2019b), obtained through an interpolation strategy that gives more importance to recent events,

$$\mathcal{R}_{\mathcal{E}}^{\text{vox}}(x,y,b) = \sum_{i=1}^{N} p_i k_b(x-x_i) k_b(y-y_i) k_b(b-t_i^*), \quad \text{with} \quad t_i^* = (B-1)\frac{t_i - t_1}{t_N - t_1}, \tag{2.3}$$

where t_i^* are the event timestamps rescaled into $[0, B-1]$, and $k_b(a) = \max(0, 1-|a|)$ is the bilinear sampling kernel proposed by (Jaderberg et al., 2015).

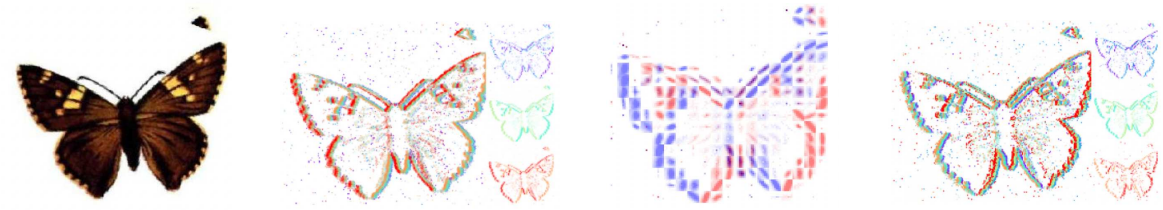
Histograms of Time Surfaces (HATS). *Histograms of Time Surfaces (HATS)* (Sironi et al., 2018b) present a dual-channel representation that advances the concept of time surfaces with a robust memory mechanism for noise mitigation. HATS are constructed by segmenting the event stream into C distinct cells, each encompassing a $K \times K$ pixel area. Within each cell c, a grid of $(2\rho+1) \times (2\rho+1)$ histograms $h_{c,p}$ is built, one for each polarity p. These histograms are computed by aggregating time surfaces $T_{e_i}(p)$, defined as:

$$T_{e_i}(p) = \begin{cases} \sum_{j \in N_{e_i}(p)} e^{-\frac{t_j - t_i}{\tau}} & \text{if } p_i = p, \\ 0 & \text{otherwise,} \end{cases}$$

where $N_{e_i}(p)$ is the cell's memory providing the set of events preceding e_i in a $[-\rho, \rho]$ spatial neighborhood. The resulting two-channel representation is an combination of normalized time surface histograms, ordered by the originating cells' locations:

$$R_{\mathcal{E}}^{HATS} = \{h_{c_j,p}\}_{j=1}^{C}, \quad h_{c_j,p} = \frac{1}{|c_j|} \sum_{e_i \in c_j} T_{e_i}(p).$$

The ρ parameter is often such that $2\rho + 1 < K$, thus reducing the initial grid resolution. Temporal resolution is also lost, as the entire temporal window is condensed into a single frame with no bins retaining temporal resolution. For these reasons, other event representations are usually preferred in deep learning applications. An example of a HATS representation is shown in Figure 2.12c.



(a) RGB (b) Voxel Grid (c) HATS (d) EST

Fig. 2.12 **Event representations.** Comparison between different event representations using the last 100ms (third saccade) of the *butterfly_0006 N-Caltech101* [116] sample. *Image taken from https://hdl.handle.net/10589/187047*

Event Spike Tensor (EST). The *Event Spike Tensor (EST)* was proposed by Gehrig et al. (Gehrig et al., 2019b) as the first end-to-end trainable grid-like representation. Unlike a voxel-grid image, EST utilizes event timestamps as pixel features, with the significance of each event's contribution being determined by a Multi Layer Perceptron (MLP) network, rather than being pre-set. Events are categorized by polarity to form a dual-channel representation for each bin. EST significantly enhances event representation by incorporating components within the transformation that can be learned, thereby automating the tuning process for a specific task. It has been successfully applied to object recognition (Gehrig et al., 2019b, 2020), optical flow prediction (Gehrig et al., 2019b), and semantic segmentation (Gehrig et al., 2020). An example of an EST representation is shown in Figure 2.12d.

2.3.3 Datasets and Simulators

Event-based cameras, being a relatively novel technology, initially had a scarcity of associated datasets. This trend is rapidly changing as the number of event-based vision datasets has seen a significant rise. These datasets can be broadly categorized into two types: those designed for motion estimation or image reconstruction (regression) tasks, and those introduced for recognition (classification) tasks. The first type includes datasets essential for developing algorithms for optical flow estimation (Rueckauer and Delbruck, 2016; Zhu et al., 2018), Simultaneous Localization and Mapping (SLAM) (Barranco et al., 2016; Delmerico et al., 2019; Weikersdorfer et al., 2014), object tracking (Hu et al., 2016), and segmentation (Alonso and Murillo, 2019; Diehl et al., 2015; O'Connor et al., 2013). The second type involves datasets tailored for recognizing objects and actions. Our focus within this section is on the datasets for object and action classification, as they are most pertinent to our research objectives. Despite the growing number of event-based datasets, as exemplified by the recent introduction of N-ImageNet (Kim et al., 2021c), there remains a significant gap when

Dataset	Task	Acquisition	# Classes	# Labels	Total time (h)
Poker-DVS [128]	Classification	real-world, still cam	4	131	2.1 sec
N-MNIST [116]	Classification	LCD, still image, moving cam	10	70,000	5.83
MNIST-DVS [71]	Classification	LCD, moving image, still cam	10	30,000	16.67
CIFAR10-DVS [129]	Classification	LCD, moving image, still cam	10	10,000	3.33
N-Caltech101 [116]	Classification	LCD, still image, moving cam	101	9,146	0.76
DVS-Caltech256 [130]	Classification	LCD, moving image, still cam	257	30,607	8.58
N-Cars [85]	Classification	real-world, moving cam	2	24,029	0.68
N-ImageNet [112]	Classification	LCD, still image, moving cam	1,000	1,781,167	24.74
N-ROD [4]	Classification	LCD, still image, moving cam	51	41,877	3.49
ASL-DVS [89]	Gesture Recog.	real-world, still cam	24	100,800	2.80
DVS128 Gesture [133]	Gesture Recog.	real-world, still cam	11	1,342	2.24
DVS-UCF-50 [130]	Action Recog.	LCD, moving image, still cam	50	6,676	13.81

Table 2.4 **Event-based datasets.** Comparison between available datasets for classification, gesture and action recognition, detection, optical flow prediction, and segmentation.

compared to the abundance of standard image-based datasets. To overcome these challenges, simulation and unsupervised learning stand as the primary strategies for training deep neural networks in absence of large-scale event-based vision tasks.

In the following, we delve into the existing event-based datasets for object and action classification tasks. Next, we explore how event-based data simulators function and contribute to this field of study.

Datasets

We summarize the main object and action classification datasets in Table 2.4.

The *Poker-DVS* datasets (Serrano-Gotarredona and Linares-Barranco, 2015) is one of the very first classification datasets to be introduced and is obtained by first quick browsing (Pérez-Carrasco et al., 2013) a deck in front of an event camera and then extracting motion-compensated pictures with a tracking algorithm. (Serrano-Gotarredona and Linares-Barranco, 2015) proposed to convert existing image-based datasets into their event-based version by artificially moving image samples on an LCD monitor and recording them with an event-based camera. The *MNIST-DVS* dataset (Pérez-Carrasco et al., 2013) is an event-based derived version of the well-known MNIST dataset (LeCun et al., 1998). It comprises 10,000 samples across 10 digit classes, with each digit captured at three distinct resolutions. These recordings, made by displaying moving images on a monitor, exhibit non-continuous motion due to the dependence on the refresh rate of the LCD screen. To address this, (Orchard et al., 2015b) introduced an alternative approach by fixing the image position and instead moving the camera. Mimicking the saccadic eye movements found in humans, they utilized a pan-tilt mechanism to create the *N-MNIST* (Orchard et al., 2015b) and the *N-Caltech101* (Orchard et al., 2015b) datasets, with the latter being a conversion of the Caltech101 (Fei-Fei et al.,

2006) dataset. N-MNIST preserves the original dataset's structure, with 60,000 training and 10,000 test samples, each at a resolution of 34×34 pixels. N-Caltech101 consists of 9,146 variable-sized images distributed across 101 categories. Subsequent conversions of well-established image datasets followed similar methodologies. The *CIFAR10-DVS* dataset by Li et al. (Li et al., 2017b) converts the CIFAR-10 benchmark (Krizhevsky, 2009) into 10,000 samples of 128×128 resolution for event-based vision. (Hu et al., 2016) transformed several frame-based collections, including the Caltech-256 (Griffin et al., 2007), with its 30,607 images across 257 classes, and the UCF-50 (Reddy and Shah, 2013) Action Recognition Dataset, which contains 6,676 samples categorised into 50 action classes, averaging 6.64 seconds each. More recently, (Kim et al., 2021c) presented the *N-ImageNet*, an event camera adaptation of the large-scale ImageNet (Deng et al., 2009). N-ImageNet is currently the most extensive object recognition dataset for event-based vision in terms of class and sample size. It includes 1,781,167 event recordings distributed over 1,000 classes, with each recording lasting 50ms.

To avoid artifacts due to their conversion procedures, researches have also introduced more realistic datasets. For instance, the *N-Cars* dataset (Sironi et al., 2018b) comprises urban scene recordings, each lasting 100ms, and categorises objects into cars and urban backgrounds. The *ASL-DVS* dataset (Bi et al., 2019), contains recordings of handshapes for American Sign Language (ASL) classification, offering 24 classes that represent ASL letters (A to Z, excluding J). Each class has 4,200 samples, approximately 100ms in duration, captured under natural conditions. The *DVS-128 Gesture* dataset (Amir et al., 2017) presents the first benchmark for gesture recognition with event-based sensors. It features 1,342 samples of 11 unique hand and arm gestures, recorded from 29 individuals across 122 trials under three different lighting scenarios.

In this thesis, we follow the procedure outlined in (Orchard et al., 2015b) to create N-EPIC-Kitchens, the first event-based dataset from an egocentric perspective (Section 5.2), and N-ROD, the first dataset designed to study the Synthetic-to-Real domain shift in event-based data (Section 5.3.3).

Simulators

Although several event-based datasets are being proposed, there is still a lack of available large-scale datasets to unlock the potential of event-based data. To address this problem, event simulators have been proposed that emulate DVS sensors' output. These systems operate by analyzing a video stream and monitoring the logarithmic brightness at each pixel.

A new ON/OFF event is triggered whenever there is a change in brightness at a pixel that exceeds a pre-determined threshold compared to the last recorded event at that location. A common practice among these methods involves capturing video at extremely high frame rates to accurately capture the visual signals. (Rebecq et al., 2018) proposed the ESIM simulator, which enhanced these techniques by adaptively adjusting the video's frame rate, thereby minimizing the need for processing and evaluating numerous frames. Additionally, Vid2E (Gehrig et al., 2020) transforms standard $30-60fps$ videos by using a slow-motion technique to interpolate frames at a variable frame rate before simulating events.

In this thesis, we conduct an in-depth analysis of the impact of the gap between simulated and real data (Sim-to-Real domain shift) on event-based data (Section 5.3). After ensuring this domain gap can be overcome with standard domain adaptation techniques, we use simulation to introduce two new datasets to unlock the potential of event-based data, namely N-EPIC-Kitchens (Section 5.2) and N-ROD (Section 5.3.3).

Chapter 3

Multi-Modal Relative Norm Alignment for Tackling the Domain Shift

A well-known problem in the literature is the so-called "domain shift", i.e., a model trained on a labeled source dataset does not generalize well to an unseen target dataset that comes from a different distribution than the source. Recent egocentric video understanding models use information from multiple modalities, such as complementary audio-visual (Zhu et al., 2021) and appearance-motion information (Munro and Damen, 2020a; Ng et al., 2018; Sevilla-Lara et al., 2019; Sun et al., 2018b), to improve accuracy and generalization performance. Despite its advantages, Multi-Modal Learning (MML) also comes with some challenges. These include figuring out how to summarize data while preserving its complementary information (Wang et al., 2020a) and understanding how to effectively combine information from multiple modalities for accurate predictions (Baltrušaitis et al., 2019). Moreover, different modalities may be impacted differently by domain shift (Lv et al., 2021), making it even more challenging to learn from them in cross-domain scenarios.

In this chapter, we propose a method to address both the cross-modal and cross-domain challenges in MML. We propose a simple loss called *Relative Norm Alignment* (RNA) loss which attempts to align the average feature norms of the different modalities to a common value. Through extensive experiments on multiple modalities (RGB, audio, optical flow), we show that this loss leads to successful generalization across domains.

The work presented in this chapter led to three publications:

- Planamente, M., Plizzari, C., Peirone, S. A., Caputo, B., & Bottino, A. (2024). Relative Norm Alignment for Tackling Domain Shift in Deep Multi-modal Classification.

International Journal of Computer Vision, 1-21.
Online Resources: [Paper]

- Planamente*, M., Plizzari*, C., Alberti, E., & Caputo, B. (2022). Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 1807-1818).*
  Online Resources: [Paper]

- Plizzari*, C., Planamente*, M., Alberti, E., Caputo, B., PoliTO-IIT Submission to the EPIC-KITCHENS-100 Unsupervised Domain Adaptation Challenge for Action Recognition.
  Third Place at the EPIC-Kitchens Unsupervised Domain Adaptation Challenge at CVPR 2021. (technical report)
  Online Resources: [Paper]

3.1 Introduction

In egocentric vision, the recording device is worn by the observer and moves with her. As a result, there are significantly more variations in lighting, perspective, and surroundings than with a fixed third-person camera. Despite numerous publications in the field, egocentric action recognition still has a major unresolved flaw known as "environmental bias" (Torralba and Efros, 2011). This problem arises from the network's heavy reliance on the environment in which the activities are recorded, which hinders the network's ability to recognize the same actions when they are performed in unfamiliar (unseen) surroundings. To illustrate its impact, we show in Figure 3.1 the relative drop in model performance from the seen to the unseen test set for the top-3 methods of the 2019 and 2020 EPIC-KITCHENS challenges (Damen et al., 2020). Generally, this issue is referred to in the literature as "domain shift", meaning that a model trained on a source labeled dataset cannot generalize well to an unseen dataset, referred to as the target. Several studies have addressed this issue by framing it as an Unsupervised Domain Adaptation (UDA) setting, where an unlabeled set of samples from the target domain is available during training (Munro and Damen, 2020b). However, the UDA scenario is not always realistic because the target domain might not be known in advance, or accessing target data at training time might be costly or simply impossible.

We argue that the true challenge lies in learning a representation that can generalize to any unseen domain, regardless of the ability to access target data during training. This approach is most commonly known as the Domain Generalization (DG) setting. Inspired by the concept of exploiting the multi-modal nature of videos (Kazakos et al., 2019a; Munro and Damen, 2020b), we utilize multi-sensory information to tackle the challenges inherent in this setting.

Although multiple modalities could potentially offer additional information, the capability of CNNs to effectively extract useful knowledge from them is somehow limited (Alamri et al., 2019; Goyal et al., 2017; Poliak et al., 2018; Wang et al., 2020a; Weston et al., 2011). We identified that a significant challenge in learning from multiple modalities stems from the tendency to prioritize one modality over others during training. To address this imbalance, we introduce a straightforward technique known as the *Relative Norm Alignment* (RNA) loss. In the context of Domain Generalization (DG) — where the model is trained without access to the target data — this loss aims to equalize the average norms of different modalities, facilitating a more balanced learning process. This ultimately leads to better generalization on unseen domains.

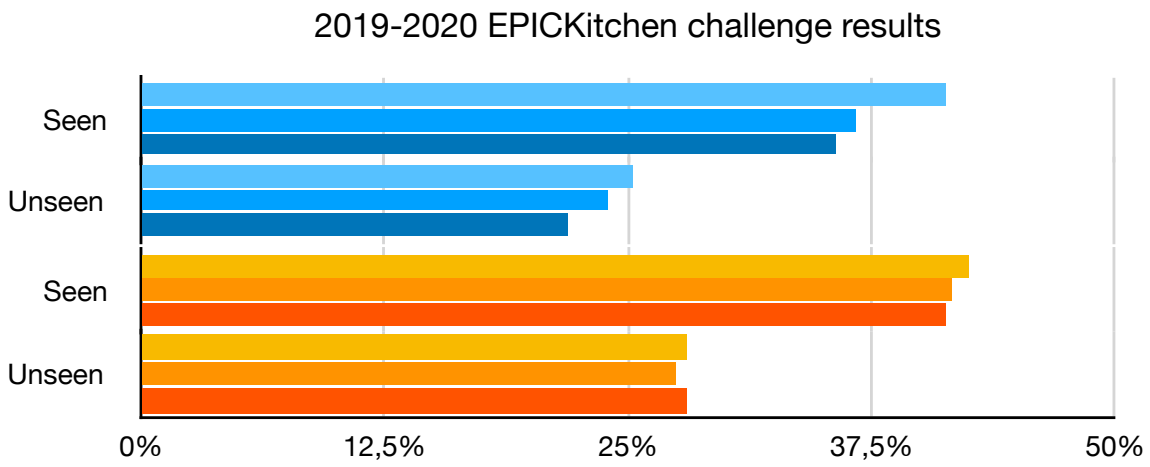


Fig. 3.1 **Seen *vs* unseen performance.** Top-3 results of the 2019 (Damen et al., 2019) and 2020 (Damen et al., 2020) EPIC-KITCHENS challenges, when testing on "Seen" and "Unseen" kitchens.

We then extend the loss to operate in the traditional UDA setting. Under the UDA setting, RNA is defined as the sum of two domain-specific terms that aim to achieve a cross-modality norm balance on both source and target domains. To further push the network to focus on features that are more transferable between domains (Xu et al., 2019a), we then extend the loss to re-balance the feature norms across domains independently for each modality. Additionally, we include in the definition of RNA an additional component to enforce similar feature norms between classes intra- and inter-domain, which ultimately helps to improve overall accuracy.

In summary, the main contributions of this section are the following:

- we bring to light the "imbalance" problem that emerges when training multi-modal networks, which leads to the network "favoring" one modality over the others during training, thereby limiting its ability to generalize (Section 3.2.1).
- we propose a new multi-modal loss, the Relative Norm Alignment (RNA) loss, designed to progressively align the relative feature norms of multiple modalities during training, thereby resulting in domain-invariant features (Section 3.2).
- we present an extensive analysis and ablation of our approach in both DG and UDA settings, showing state-of-the-art or competitive performances on all benchmarks (Section 3.3).

3.2 RNA: Relative Norm Alignment

Next, we describe some preliminary intuitions and motivations behind the proposed approach in Section 3.2.1. We then describe the proposed Relative Norm Alignment (RNA) loss, designed to reduce domain shift in Multi-Modal Learning (MML) by aligning the average feature norms across different modalities (cross-modal alignment) and different domains (cross-domain alignment), both globally and at the class level. We adapt RNA to operate both in the Domain Generalization (DG) setting and the Unsupervised Domain Adaptation (UDA) setting. We detail RNA implementation in both settings in the following.

3.2.1 Intuition and motivation

A widely adopted method for addressing the task of first-person action recognition in research is the utilization of multi-modal approaches (Cartas et al., 2019; Kazakos et al., 2019a, 2021a; Lin et al., 2019; Munro and Damen, 2020b; Wang et al., 2016). Despite the richer information multi-modal systems offer compared to single-modal ones, their advantages in performance are often marginal and inconsistent (Alamri et al., 2019; Goyal et al., 2017; Poliak et al., 2018; Wang et al., 2020a; Weston et al., 2011). The issue of limited performance gains has been linked to overfitting by (Wang et al., 2020a), who proposed mitigating this by adjusting the loss value for each input type using distinct hyperparameters. However, this solution requires an intricate step of fine-tuning that is heavily dependent on both the specific task and the dataset used. We propose to tackle the challenges associated with multi-modal learning from an alternative perspective.

Norm *imbalance*. We hypothesize that an imbalance between different modalities during training hinders the network's ability to learn from them equally. This theory is supported by the observation that the hyperparameters identified in (Wang et al., 2020a) vary considerably based on the modality. To verify this theory, we conducted a simple experiment on RGB and audio data, whose results are shown in Figure 3.2-a. Independently trained RGB and audio streams perform comparably well during testing. Yet, when trained jointly but tested separately, RGB's performance drops in comparison to audio's, indicating that multi-modal training negatively affects RGB stream optimization.

This led us to consider whether the imbalance observed between modalities during training could also be present in a multi-source scenario. *Could one source disproportionately influence another, thereby diminishing the overall model's effectiveness?* With these questions

in mind, we sought a method to measure the information each modality's final embedding carries, hoping to shed light on the reason for such an imbalance.

The *mean feature norms*. Several works highlighted that there exists a strong correlation between the mean feature norms and the amount of "valuable" information for classification (Ranjan et al., 2017; Wang et al., 2017a; Zheng et al., 2018). Notably, cross-entropy loss tends to favor features that are well-differentiated and possess high norm values, as noted by (Wang et al., 2017a). Additionally, the principle proposed by (Ye et al., 2018) suggests that a modality's representation is less informative during inference if it has a lower norm, which is summarized as the Smaller-Norm-Less-Informative hypothesis. Taken together, these findings indicate that the $L2$-norm of features can reflect their informational value, serving as a useful metric for detecting imbalances between training modalities. Our analysis of feature norms revealed that the average norms of audio features (approximately 32) were significantly higher than those for RGB (approximately 10) in the training set. This disparity is reflected on the test set, as shown in Figure 3.2 on the left, where the modality with the lower norm exhibits lower performance.

Motivated by these results, we propose a simple but effective loss whose goal is to re-balance the mean feature norms during training across multiple sources, so that the network can fully leverage the benefits of joint training, particularly in scenarios involving cross-domain data. The process of norm re-balancing results in improved performance for both modalities, as demonstrated in Figure 3.2, right. It is important to clarify that using the concept of smaller norms being less informative primarily highlights the network's bias towards the audio modality due to its higher norm relative to RGB. However, this observation does not mean that RGB is inherently less valuable for the task. After the norms are re-balanced, their range more closely approximates that of the original RGB norms, indicating that both modalities are equally important.

3.2.2 Relative Norm Alignment loss

Problem Definition. Let us consider data $\mathcal{X}_{\mathcal{S}} = \{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$ from a source distribution $\mathcal{S}$, where n_s represents the total number of samples, and each sample $x_{s,i}$ is associated with a label $y_{s,i}$ from the label space $\mathcal{Y}_s$. Each sample $x_{s,i}$ contains multiple modalities, i.e., $x_{s,i} = \{x_{s,i}^1, \ldots, x_{s,i}^M\}$, where $x_{s,i}^m$ indicates the m^{th} modality of the i^{th} sample and M is the number of modalities. The target domain $\mathcal{T}$ includes n_t annotated target samples $\mathcal{X}_{\mathcal{T}} = \{x_{t,i}\}_{i=1}^{n_t}$, each characterized by the same M modalities of the source samples, i.e., $x_{t,i} = \{x_{t,i}^1, \ldots, x_{t,i}^M\}$.

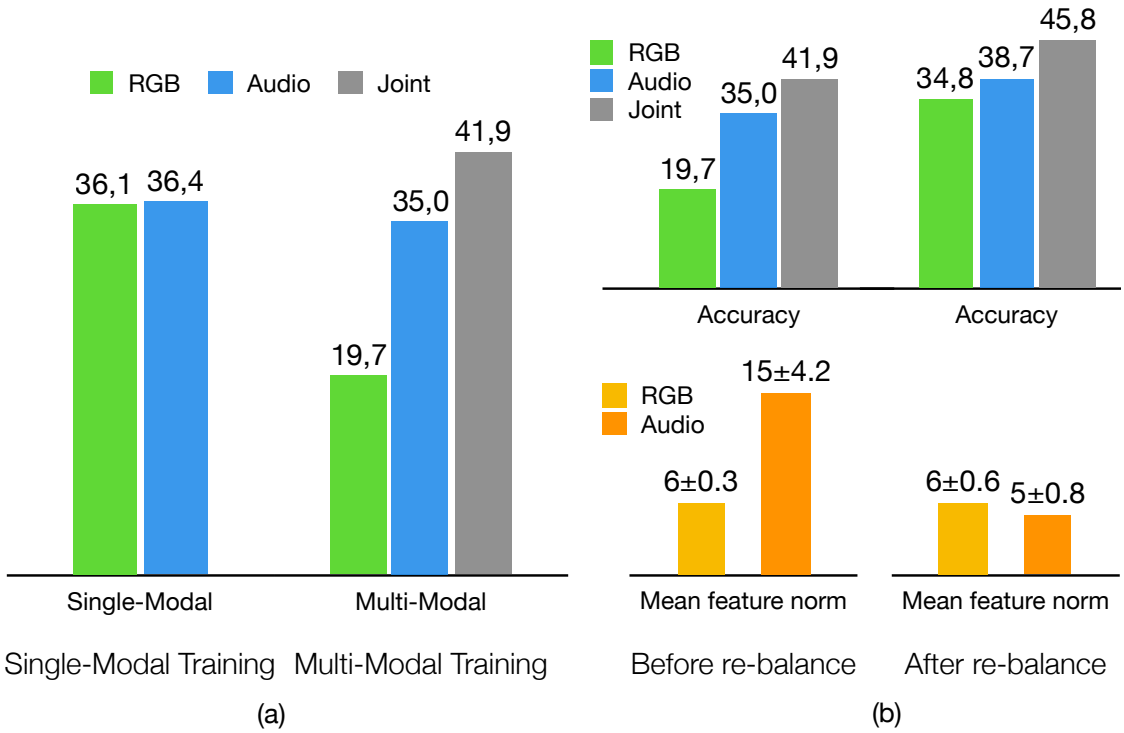


Fig. 3.2 **Norm imbalance.** By jointly training, and testing on separate streams, the RGB performance drop (a). "imbalance" at feature-norm level which, when mitigated, leads to better performance (b).

We assume that the distributions of all domains involved are distinct, denoted as $\mathcal{D}^j_{d_1} \neq \mathcal{D}^k_{d_2}$, where d_1 and d_2 refer to the domains (source or target) and j and k indicate different modalities within the same domain or the same or different modalities across different domains. We assume that the label space $\mathcal{Y}s$ is identical to $\mathcal{Y}t$, indicating a shared label space between the source and target domains.

Relative Norm Alignment loss. In the following, we consider for simplicity a single-source single-target setting in which only two modalities are available. In Section 3.2.2 we detail how the approach can be extended to work with any number of modalities.

Each input sample is denoted as $x_i = (x_i^u, x_i^v)$, where v and a represent the two modalities, e.g., visual and audio modality. As illustrated in Figure 3.3, each input modality m is passed through a dedicated feature extractor F^m. The resulting features $f^m = F^m(x_i^m)$ are then processed by a classifier G^m, which produces score predictions for the m^{th} modality of the i^{th} sample. Finally, the prediction scores from all modalities are combined using a late fusion approach to derive the final classification. It is important to note that in UDA settings, the F^m feature extractors are shared between the source and target domains.

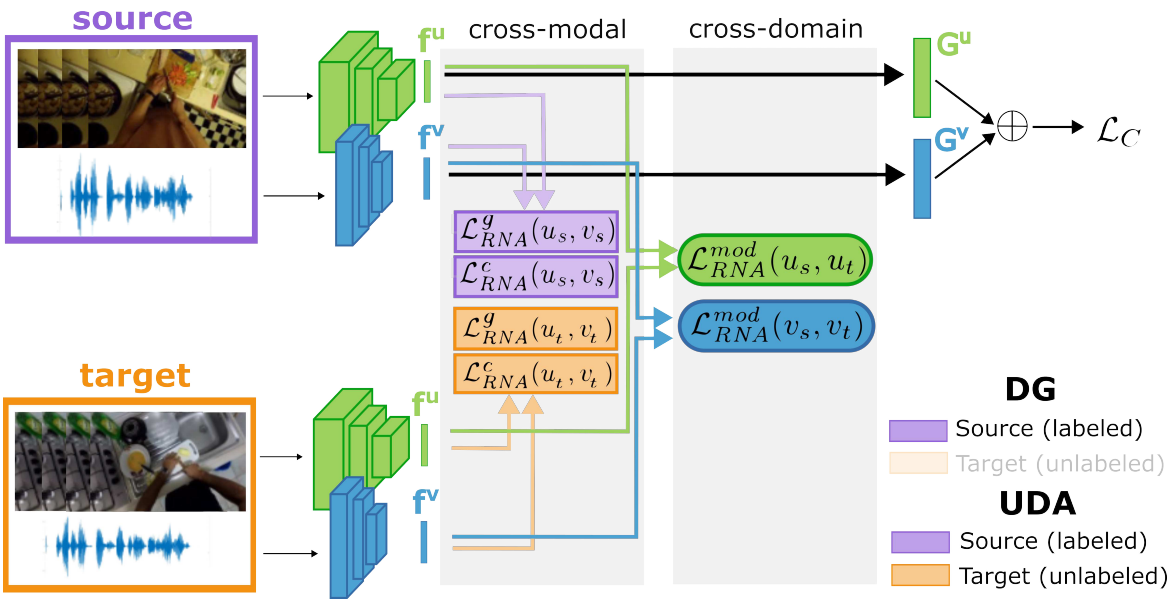


Fig. 3.3 **Method architecture.** Labeled source samples and unlabeled target samples from modalities u (e.g., visual) and v (e.g., audio) are fed to their corresponding feature extractors. $\mathcal{L}_{RNA}$ is designed to maintain a balance between the relative feature norms of the two modalities, achieved through a combination of domain-specific cross-modal components ($\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$) and cross-domain components ($\mathcal{L}^{mod}_{RNA}$) for each f_u and f_v modality feature. In Domain Generalization, only the components computed on the source domain are utilized. Finally, a classification loss $\mathcal{L}_C$ is applied on the output of the modality classifiers G_u and G_v.

The main idea behind the proposed loss is the concept of *mean feature norm distance*. We denote with $h(x_i^m) = (\|\cdot\|_2 \circ f^m)(x_i^m)$ the L_2-norm of the features f^m of the m-th modality, and compute the *mean-feature-norm distance* (δ) between the two modality norms f^u and f^v as

$$\delta(h(x_i^u), h(x_i^v)) = |\mathbb{E}[h(x_i^u)] - \mathbb{E}[h(x_i^v)]| \tag{3.1}$$

where $\mathbb{E}[h(x_i^m)]$ corresponds to the mean features norm for each modality. Figure 3.4 illustrates the norm $h(x_i^u)$ of the i-th visual sample and $h(x_i^v)$ of the i-th audio sample, by means of segments of different lengths arranged in a radial pattern. The mean feature norm of the k-th modality is represented by the radius of the two circumferences, and δ is represented as their difference. The goal is to minimize the δ distance through a loss function that aims to align the mean feature norms of the two modalities. In other words, this means constraining the features from both modalities to reside on a hypersphere with a predetermined radius.

We propose a Relative Norm Alignment (RNA) loss, which is defined as:

$$\mathcal{L}^g_{RNA}(u, v) = \lambda_g \left(\frac{\mathbb{E}[h(X^u)]}{\mathbb{E}[h(X^v)]} - \frac{\mathbb{E}[h(X^u)]}{\mathbb{E}[h(X^v)]} \right)^2 \tag{3.2}$$

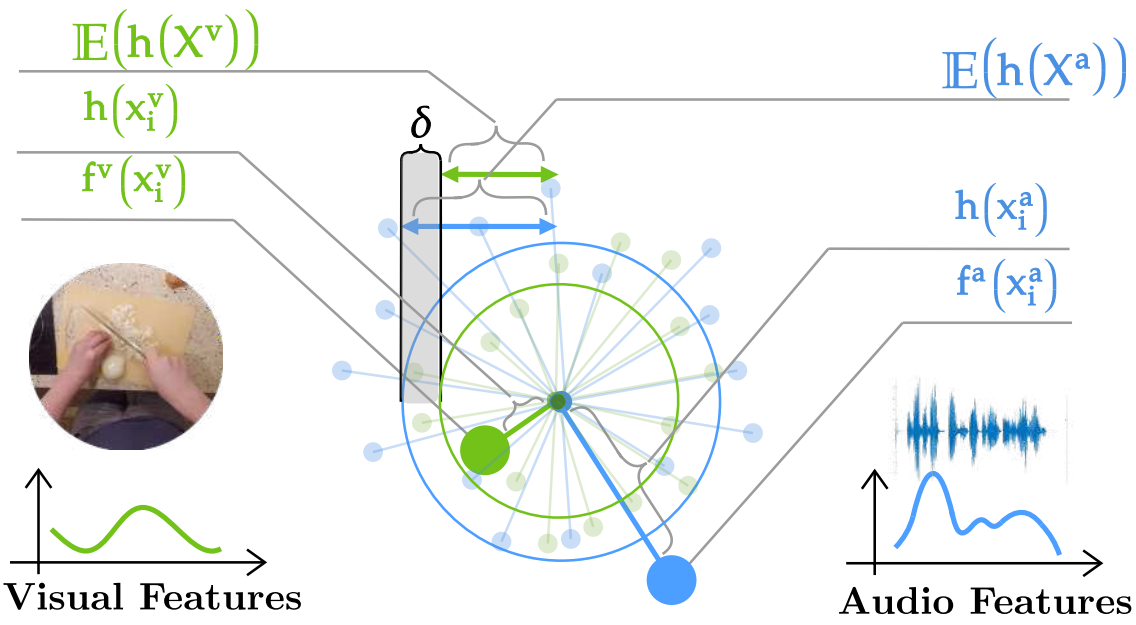


Fig. 3.4 **Relative norm alignment.** The norm $h(x_i^v)$ of the i-th visual sample (left) and $h(x_i^a)$ of the i-th audio sample (right) are represented by segments of different lengths.The radius of the two circles represents the mean feature norm of the two modalities, and δ signifies their discrepancy. By minimizing δ, we encourage the audio and visual feature norms to align.

where $h(x_i^m) = (\|\cdot\|_2 \circ f^m)(x_i^m)$ is the L_2-norm of m^{th} modality features of the i^{th} sample, $\mathbb{E}[h(X^m)] = 1/B \sum_{x_i^m \in X^m} h(x_i^m)$ is the mean feature norm for the m^{th} modality, computed over the B samples composing the batch, and λ_g weights $\mathcal{L}_{RNA}^g$. This dividend/divisor structure is designed to promote an alignment of the norms between the two modalities, aiming for an *optimal balance* between the embeddings of the two. Additionally, squaring the difference pushes the network to make more substantial adjustments when the ratio of the norms of the two modalities significantly deviates from unity, thereby accelerating convergence.

Conceptually, aligning the norms of the two modalities is similar to applying a "strict" constraint that matches them to a fixed value k. This approach, termed *Hard Norm Alignment* (HNA), is encapsulated in the $\mathcal{L}_{HNA}$ loss formula:

$$\mathcal{L}_{HNA} = \sum_m (\mathbb{E}[h(X^m)] - k)^2, \tag{3.3}$$

where k is the same across all modalities. However, our $\mathcal{L}_{RNA}$ formulation effectively reduces the gap between the norms of the two distributions without the need for the extra k hyperparameter. Opting for a subtraction approach ($\mathcal{L}_{RNA}^{sub}$) to directly minimize δ^2 (Equation 3.1) presents a simpler and viable alternative. The choice for this method stems from the consideration that a significant difference between k and the expected norms of the

modalities could result in a high loss value, necessitating precise adjustment of the weights and thereby heightening the network's sensitivity to these weights (Kendall et al., 2018). This dividend/divisor arrangement guarantees the loss remains within the range $(0, 1]$.

The main goal of the RNA loss is to teach the model how to effectively utilize the correlation between multiple modalities' norms at feature level to develop a robust and general classification model. This focus on *feature-level* adjustment is key to achieving generalization performance, distinguishing our approach from simple input-level normalization or pre-processing strategies. Importantly, normalization at the input stage may not align well with pre-trained models and is impractical in Domain Generalization (DG) settings where target data is inaccessible during training. This means there is no access to the target distribution, and each domain might necessitate a unique normalization approach. Additionally, *learning to re-balance* norms instead of applying conventional projection methods for feature normalization is driven by two considerations. Firstly, by integrating feature normalization within the learning process through model weights, the network is better equipped to address the "norm imbalance" issue not just during training but also in inference, adapting to a normalized feature space learned during training. Secondly, explicit normalization techniques like batch normalization adjust each feature element to a scaled normal distribution independently, which does not guarantee the alignment of the overall mean feature norms across modalities.

The overall architecture is finally trained by minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}^g_{RNA} \tag{3.4}$$

where $\mathcal{L}_C$ is the standard *cross-entropy loss* on source data.

Per-class alignment. The formulation in Eq. 3.2 has a limitation: the *global* cross-modal alignment facilitated by $\mathcal{L}^g_{RNA}$ can result in imbalanced norms between modalities at the class level. This imbalance may lead to a preference for one modality over others when classifying specific classes. To mitigate this issue, we propose the following enhancements to the RNA framework.

First, we introduce an intra-domain class constraint, $\mathcal{L}^c_{RNA}$, to rectify the cross-modal norm imbalance at the class level. It is defined as follows:

$$\mathcal{L}^c_{RNA}(u, v) = \lambda_c \sum_{c=1}^{C} \left(\frac{\mathbb{E}[h(X^u_c)]}{\mathbb{E}[h(X^v_c)]} - \frac{\mathbb{E}[h(X^v_c)]}{\mathbb{E}[h(X^u_c)]} \right)^2 \tag{3.5}$$

where λ_c is the weight of the loss, and $\mathbb{E}[h(X_c^m)]$ represents the average norm of the features for modality m for samples belonging to class c, with C being the total number of classes. Combining the two components we have previously defined, the extended RNA formulation in DG settings becomes:

$$\mathcal{L}_{RNA} = \mathcal{L}^g_{RNA}(u_s, v_s) + \mathcal{L}^c_{RNA}(u_s, v_s) \tag{3.6}$$

RNA for Domain Adaptation (UDA). The RNA objective in UDA can be defined as:

$$\mathcal{L}_{RNA} = \mathcal{L}^g_{RNA}(\mathcal{S}) + \mathcal{L}^g_{RNA}(\mathcal{T}) + \mathcal{L}^c_{RNA}(\mathcal{S}) + \mathcal{L}^c_{RNA}(\mathcal{T}) \tag{3.7}$$

where $\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$ are defined as the losses in Eq. 3.2 and Eq. 3.5 applied to both source and target domains. Note that for target samples, pseudo-labels are employed to categorize them into classes when calculating $\mathcal{L}^c_{RNA}$.

In the UDA setting, alignment is performed separately within each domain. Consequently, substantial discrepancies in the average feature norms between source and target domains may persist. Such variations often stem from domain-specific features that, while large in the source domain training, may exhibit reduced activations in the target domain (Barbato et al., 2021; Xu et al., 2019a). This disparity can significantly impact the model's overall accuracy.

We extended the $\mathcal{L}_{RNA}$ loss to align both the average norms and the per-class norms of features within each modality across domains. This adjustment allows the network to prioritize features with higher transferability between domains (Xu et al., 2019a). To this end, we include the following term in the RNA formulation:

$$\mathcal{L}^{mod}_{RNA}(m_s, m_t) = \mathcal{L}^g_{RNA}(m_s, m_t) + \mathcal{L}^c_{RNA}(m_s, m_t)$$

where $m \in \{u, v\}$. The resulting RNA formulation for the UDA setting is:

$$\begin{aligned} \mathcal{L}_{RNA} = & \mathcal{L}^g_{RNA}(u_s, v_s) + \mathcal{L}^g_{RNA}(u_t, v_t) + \\ & \mathcal{L}^c_{RNA}(u_s, v_s) + \mathcal{L}^c_{RNA}(u_t, v_t) + \\ & \mathcal{L}^{mod}_{RNA}(u_s, u_t) + \mathcal{L}^{mod}_{RNA}(v_s, v_t) \end{aligned} \tag{3.8}$$

The individual contribution of the three losses is exemplified in Figure 3.5. $\mathcal{L}^g_{RNA}$ is responsible for the global alignment of modality norms within each domain. $\mathcal{L}^c_{RNA}$ ensures the alignment of modality norms for each class within the domains. $\mathcal{L}^{mod}_{RNA}$ focuses on aligning the norms between domains for each modality independently. In the DG setting, $\mathcal{L}^c_{RNA}$ enhances

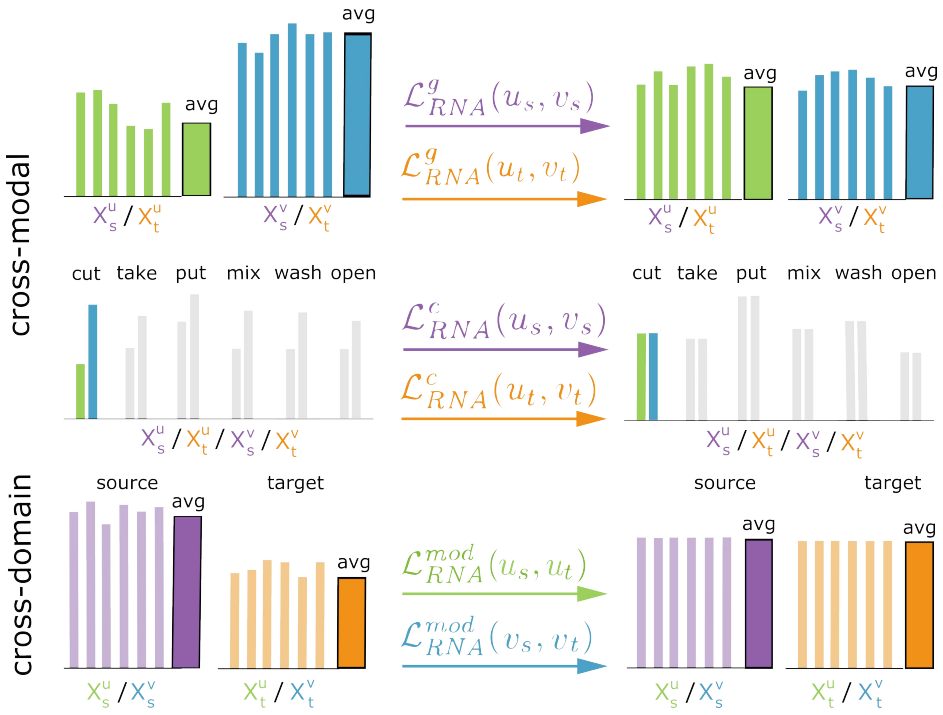


Fig. 3.5 ***Distinct impacts* of $\mathcal{L}_{RNA}$ components on feature norms.** For each plot, norms for every class within a given modality and domain are displayed (u or v, associated with source or target). **First row:** $\mathcal{L}^{g}_{RNA}$ is designed to reduce the overall average norms (indicated by the expanded bars on the right) for modalities u and v. **Second row:** $\mathcal{L}^{c}_{RNA}$ is focused on ensuring norms are even at the class level. **Third row:** $\mathcal{L}^{mod}_{RNA}$ aims at re-balancing class and average norms for the same modality across domains. Each diagram illustrates the norms before (on the left) and after (on the right) the implementation of the specific $\mathcal{L}_{RNA}$ component.

the effectiveness of $\mathcal{L}^{g}_{RNA}$ by ensuring that norms are aligned per class to a unified standard. The integration of $\mathcal{L}^{mod}_{RNA}$ within UDA complements the other two losses by aligning the average and per-class norms of modalities between the source and target domains.

Extension to more than two modalities. The RNA objective in Eqs. 3.6 and 3.8 can be extended to more than two modalities. In DG, the loss can be rewritten as:

$$\mathcal{L}_{RNA} = \mathcal{L}_{RNA}(\mathcal{S}) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \mathcal{L}_{RNA}(i_s, j_s) \tag{3.9}$$

where i and j span the M modalities. Similarly, the UDA loss becomes:

$$\mathcal{L}_{RNA} = \mathcal{L}_{RNA}(\mathcal{S}) + \mathcal{L}_{RNA}(\mathcal{T}) + \sum_{i=1}^{M} \mathcal{L}^{mod}_{RNA}(i_s, i_t)$$

where $\mathcal{L}_{RNA}(\mathcal{S})$ and $\mathcal{L}_{RNA}(\mathcal{T})$ are the loss in Eq. 3.9 for the source and target domains, respectively.

Additional learning objectives. In addition to the loss defined in Eq. 3.8, for enhancing the domain-invariant characteristics of the features, we employ adversarial domain alignment techniques (Ganin and Lempitsky, 2015a; Wang et al., 2019b). This approach is in line with methods utilized in recent UDA studies (Chen et al., 2019; Jamal et al., 2018; Munro and Damen, 2020a; Wei et al., 2022), involving the integration of a classifier asked to distinguish whether features originate from the source or the target domain. This classifier is connected to the feature extractors through a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015a). Consequently, the domain classification loss $\mathcal{L}_d$ is scaled by a factor λ_d and incorporated into the total loss.

The loss framework we have introduced, which combines $\mathcal{L}_{RNA}$ and $\mathcal{L}_d$, is designed to enhance both the informativeness and the domain-invariant characteristics of the embeddings across different modalities. However, these components of loss influence only the feature extractors F^m and do not extend their impact through the classifier. As a consequence, the classifier is exposed solely to source data during training, lacking any engagement with the target data. This setup leads to a scenario where the classifier is optimized to merge multi-modal features effectively for increased accuracy within the source domain, while it overlooks the *classification uncertainty* that may arise with target data.

To tackle this issue in the UDA setting, a widely adopted strategy involves the application of a *mutual information criterion* (Bridle et al., 1991) to target data. This method not only aims to minimize prediction uncertainty but also encourages an even distribution of samples across classes. The technique employs an Information Maximization (IM) loss (Bridle et al., 1991), which is calculated as the difference between the average entropy of the model's predictions and the entropy of the average prediction across the target data:

$$\mathcal{L}_{IM} = -\mathbb{E}_{x\in\mathcal{X}_\mathcal{T}} \sum_{c=1}^{C} p_c(x) \log p_c(x) + \sum_{c=1}^{C} \bar{p}_c \log \bar{p}_c$$

where C is the total number of classes, p_c is the posterior probability for class c, and $\bar{p}_c$ is the mean output score for the current batch.

When integrating all components, we train the model in the UDA setting to minimize the following loss:

$$\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{RNA} + \lambda_d \mathcal{L}_d + \lambda_{IM} \mathcal{L}_{IM}$$

where $\mathcal{L}_{RNA}$ is from Eq. 3.8 and λ_{IM} is the IM loss weight.

3.3 Experiments

In this section, we aim to validate the effectiveness of our proposed approach through empirical evaluation on two multi-modal egocentric vision benchmarks: the EPIC-KITCHENS-100 (EK100) (Damen et al., 2022) and the EPIC-KITCHENS-55 (EK55) (Damen et al., 2018) datasets. The rest of the section is organized as follows. We introduce results on EK100 in Section 3.3.1 and on EK55 in Section 3.3.2. For each, we describe the baselines and evaluation protocol used, and implementation details. Finally, an ablation study is given in Section 3.3.3.

3.3.1 Experiments on EK100

Experimental Setup

In this section, we describe the evaluation protocol used the baselines, along with information about input pre-processing and implementation details.

Evaluation Protocol. We follow the experimental setup for UDA proposed in (Damen et al., 2022). The dataset consists of two splits, *source* and *target*, containing labelled and unlabelled samples respectively. In the *intra-domain* setting, the train and test sets are in the same visual domain, *i.e.* clips have been recorded in the same kitchens in both splits, whereas, in the *cross-domain* setting, the training and testing sets are from different kitchens (*location shift*) or from the same kitchens but recorded after a long temporal interval of several years (*temporal shift*). Actions are annotated with *(verb, noun)* pairs from a set of 97 verbs and 300 nouns. Models are evaluated in terms of top-1 and top-5 accuracy for verb, noun and action predictions. The latter is a combination of the verb and noun labels and is used to evaluate the ability of the network to predict both. All experiments described in this section utilize the three modalities (RGB, audio, and optical flow) provided by the dataset. This work's findings are presented based on the validation split, though prior research has similarly shown the efficacy of RNA on the test data as well (Planamente et al., 2022a; Plizzari et al., 2021).

Baselines. Our method is evaluated against MM-SADA (Munro and Damen, 2020a), TA^{3}N (Chen et al., 2019), and CIA (Yang et al., 2022a). Specifically, the MM-SADA framework is originally designed to work with RGB and optical flow modalities. To adapt the

audio modality, we adopt a two-branch strategy, creating separate pathways for RGB-Flow and RGB-Audio combinations, similar to the approach described in (Planamente et al., 2022b). An adversarial branch is then independently applied to each modality, ensuring a tailored and effective adaptation process. Finally, since our DG approach is primarily focused on improving the multi-modal learning capabilities of the model, we extend our analysis to include the Gradient Blending (GB) technique (Wang et al., 2020a) as a DG comparison.

Input. Following the procedure described in (Kazakos et al., 2019a), RGB and optical flow modalities are processed by uniformly sampling 25 frames, and the audio modality is processed by extracting segments lasting 1.28 seconds, each aligned with the action. For both training and inference, five segments from each modality are chosen and input into the network.

Implementation Details. Frame-level features $f_m \in \mathbb{R}^{25\times1024}$ for each modality m are derived from a TBN framework (Kazakos et al., 2019a), initially pre-trained on Kinetics (Kay et al., 2017) and subsequently fine-tuned for the source domain as per the approach described in (Damen et al., 2022). A selection of five frame features per segment is uniformly made and processed through a linear layer, followed by a ReLU activation function and a dropout layer with a rate of 0.5. The frame features undergo temporal integration via a TRN (Zhou et al., 2018) module, resulting in action-level features $f'_m \in \mathbb{R}^{1024}$[1]. The features are then divided into two segments, $f'_{m,v}$ and $f'_{m,n} \in \mathbb{R}^{256}$, through a linear layer, labeled as *verb features* and *noun features* respectively. These segments are subsequently directed towards two separate classifiers for generating modality-specific logits for verbs ($y_{m,v}$) and nouns ($y_{m,n}$). The training of the network for action recognition incorporates the usage of cross-entropy loss on the summed *per-modality* logits. We enhance the RNA framework by distinctly applying the alignment losses to the verb and noun features, right before the final classifier. This strategy of applying RNA losses ensures that the alignment impact is maximally proximate to the classifier, which is significantly influenced by the values of feature norm. Training extends over 30 epochs with a batch size comprising 128 samples, employing an SGD optimizer with a momentum of 0.9 and a weight decay of 10^{-4}. The initial learning rate is set to 0.003 and is decreased by a factor of 10 following the 10th and 20th epochs.

Results

Table 3.1 presents Top-1 and Top-5 classification accuracies for verbs, nouns, and actions on both DG and UDA settings. Alongside each method, we report improvements in average

[1]Until this stage, the procedure adheres closely to the official implementation provided for the EK100 UDA challenge (Damen et al., 2022).

Methods	Verb@1	Noun@1	Action@1	Verb@5	Noun@5	Action@5
			DG			
Source Only	47.14	27.35	18.99	75.27	49.36	41.82
MM-SADA (SS)	47.76	27.93	19.15 (▲ +0.16)	77.07	49.77	42.90 (▲ +1.08)
Source Only	50.27	29.04	19.96	81.74	52.14	46.74
GB	50.18	**29.60**	20.26 (▲ +0.3)	81.82	52.57	**46.86** (▲ +0.12)
Source Only	46.79	26.79	18.29	75.39	48.44	41.36
Our (DG)	**50.75**	27.92	19.81 (▲ +1.52)	80.64	51.37	45.33 (▲ +3.97)
Source Only†	49.81	28.55	19.77	81.10	51.90	46.22
Our† (DG)	50.20	29.31	**20.30** (▲ +0.53)	**81.85**	**52.68**	46.76 (▲ +0.54)
			UDA			
Source Only	46.70	27.78	19.20	75.42	48.27	42.12
TA3N	48.44	28.87	19.61 (▲ +0.41)	75.95	50.12	43.36 (▲ +1.24)
Source Only	47.14	27.35	18.99	75.27	49.36	41.82
MM-SADA	48.44	28.26	19.25 (▲ +0.26)	77.56	50.59	43.41 (▲ +1.59)
Source Only	47.69	28.48	19.61	-	-	-
CIA	48.34	**29.50**	**20.30** (▲ +0.69)	-	-	-
Source Only	46.79	26.79	18.29	75.39	48.44	41.36
Our (UDA)	**50.82**	29.19	20.05 (▲ +1.76)	**80.89**	**52.18**	**46.04** (▲ +4.68)

Table 3.1 **Results on EK-100.** Classification accuracies (%) on EK100 (Damen et al., 2022) reported in terms of Top-1 and Top-5 classification accuracy across noun, verb, and action metrics. Δ Acc. represents the average improvement in Top-1 accuracy. †These experiments employ cross-entropy loss on both the fused logits and the *per-modality* logits. The best results are highlighted in **bold**, with the runner-up in underlined.

Top-1 and Top-5 accuracies for actions relative to the respective Source Only baseline, which involves training on source domains and testing on the test set without applying any adaptation strategy.

For the DG setting, we compare our approach to two alternative methods. Firstly, we evaluate against a variant of MM-SADA (Munro and Damen, 2020a) known as MM-SADA (SS), which incorporates the self-supervised alignment task tailored for the source domain modalities, omitting the adversarial alignment element of the original approach as it necessitates target domain data. Secondly, we consider Gradient Blending (GB) (Wang et al., 2020a), a technique that attempts to find the ideal combination of modalities based on their tendency to overfit. This optimal combination is derived by integrating a dedicated cross-entropy loss for each modality with a fusion loss, all weighted appropriately[2]. When

[2]Note that the conventional GB method utilizes only RGB and Audio modalities. For this study, the optimal loss weights were adopted from (Damen et al., 2020), and the weights corresponding to the Flow component, which was absent in the original formulation, were tuned for our research purposes

analyzing accuracy results across the different categories, it can be observed that GB performs best, while our approach ranks as the runner-up and MM-SADA (SS) lags slightly behind. However, when considering the improvements relative to the Source Only baseline, our method shows higher improvement on the action category compared to GB (+1.52% and +3.97% vs +0.3% and +0.12%). This result suggests that our method contributes more significantly to reducing domain shift. The approach proposed in (Wang et al., 2020a) bears similarities to our method in terms of enhancing the balance between modalities for improved classification accuracy. To delve deeper, we conducted further experiments by applying our method to the Source Only results achieved through Gradient Blending. Specifically, we utilized multiple classification losses without adjusting their weights. These additional experiments are denoted by a † symbol. The outcomes, presented in Table 3.4, are encouraging. In this variation, our method achieves the highest action accuracy when compared to all DG baseline methods. Notably, our standard approach tackles the alignment challenge with an adaptive strategy that, distinct from GB, does not depend on the specific model or dataset and is controlled by only two hyperparameters: λ_g and λ_c.

In the UDA experiments, we observe that our method ranks second in terms of Top-1 noun and action accuracy, with CIA being the best performing method. However, it should be noted that CIA's evaluation starts from a higher Source Only result. On the other hand, our method achieves the best results in Top-1 verb accuracy. In terms of improvements, RNA shows significantly higher accuracy improvements compared to all other competitors in both Top-1 and Top-5 accuracies. Additionally, our performance on all the evaluation metrics aligns closely with that of other proposed baselines. Importantly, a substantial part of these improvements is evident during the DG phase, in which the target domain is not accessed. Interestingly, our best DG results demonstrates strong competitiveness and comparability with CIA, the current state-of-the-art in UDA. This underscores the generalization capability of our method in effectively handling domain shifts.

3.3.2 Experiments on EK55

Experimental Setup

In this section, we outline evaluation protocol used and the baselines, and provide information on input pre-processing and implementation details.

Evaluation protocol. We use the EPIC-KITCHENS-55 dataset (Damen et al., 2018) and we adopt the same experimental protocol of (Munro and Damen, 2020b), where the three

kitchens with the largest amount of labeled samples are handpicked from the 32 available. We refer to them here as D1, D2, and D3 respectively. We evaluate performance in both a single-source setting ($D_i \rightarrow D_j$) on these three domains. In the experiments, we restrict our analysis to the visual and motion (RGB+Flow) and visual and audio (RGB+Audio) modality combinations, which are the ones recent work in the literature focus on.

Baselines. We compare our results with several state-of-the-art UDA methods. The first group (GRL (Ganin et al., 2016), MMD (Long et al., 2015), AdaBN (Li et al., 2018d), and MCD (Saito et al., 2018)) includes approaches originally developed as image-based methods and later adapted to work with video inputs. The second group includes more recent methods such as MM-SADA (Munro and Damen, 2020a), the contrastive-based methods proposed by (Kim et al., 2021b) and STCDA (Song et al., 2021b), and the recently published CIA (Yang et al., 2022a). In our comparison, we use the results reported in the original paper for each baseline.

Input. For our study, we utilized various sampling methods to ensure a fair comparison with previous work. With *dense sampling*, we randomly chose a series of 16 consecutive frames from each video. For *uniform sampling*, we selected 16 frames distributed evenly across the video. During testing, we followed the training sampling method but used five clips instead of one, averaging the results as per the suggestion in (Wang et al., 2016). Following the experimental setup from (Munro and Damen, 2020a), we applied random cropping, scale adjustments, and horizontal flips to enrich our training data. During testing, we only used central cropping. For audio data, as described by (Kazakos et al., 2019a), we converted the audio track into a 256×256 matrix that captures the log spectrogram. We first extracted the audio from the video, sampled it at 24kHz, and then processed it using the Short-Time Fourier Transform (STFT) with a 10ms window length, a 5ms step size, and 256 frequency bands. The same sampling strategy used for RGB was also applied to optical flow inputs.

Implementation details. In our setup, both the RGB and Flow streams employ the I3D model, pre-trained on the Kinetics dataset (Kay et al., 2017), in line with the experimental framework of (Munro and Damen, 2020a). For audio feature extraction, we utilize the BN-Inception model pre-trained on ImageNet, as detailed by (Kazakos et al., 2019a). These feature extraction models are trained from start to finish. Each modality m generates features represented as $f_m \in \mathbb{R}^{1024}$. We compute logits for each modality using a separate linear layer, which are then combined. The network is trained over 5000 iterations with the SGD optimizer, momentum set at 0.9, and a weight decay of 10^{-7}. For RGB and Flow, the learning rate starts at 0.001, reducing to 2×10^{-4} after 3000 steps. For Audio, the starting learning

rate is 0.001, which is decreased tenfold at the 1000, 2000, and 3000 step marks. We set the batch size at 128.

Results

We report results on both DG and UDA settings in Table 3.2. We categorise the results based on the sampling approach for each modality: dense (D) or uniform (U). The majority of the baseline methods utilize dense sampling (D-D), with CIA being the only method using uniform sampling (U-U) for both modalities. Our findings reveal that uniform sampling, as utilized by CIA, surpasses methods based on dense sampling, supporting the insight from (Chen et al., 2021) that uniform sampling generally provides better results. Our UDA strategy outperforms all existing methods for both sampling types, improving by 0.5% on D-D sampling and 2.2% on U-U sampling. We also explored a hybrid sampling approach (D for RGB and U for Flow). Interestingly, using this sampling the Source Only method demonstrates impressive results. Since none of the baselines use this sampling, we only present our results for the Source Only, DG, and UDA. We note that the Source Only method already achieves remarkable results (up to 3% better than our method with uniform sampling). This improvement might be attributed to the mixed sampling's capacity to better leverage the unique characteristics of each modality: dense sampling captures finer static details in RGB, while uniform sampling across a broader temporal span enriches the dynamic Flow information. Our approach improves over the Source Only baseline by 1.37% and 2.38% in the DG and UDA setting respectively.

When integrating RGB with Audio, results are slightly inferior than the RGB+Flow combination. This observation suggests that audio information serves as a less informative modality compared to optical flow in this context. Our DG models improves by up to 4.37% over the Source Only, while our UDA model achieves the best results (7% improvement over Source Only and 1% improvement over the state-of-the-art method). Moreover, the performance in the DG context is on-par with that in the UDA scenario, showing a slight difference of -1.01% for RGB+Flow and -2.55% for RGB+Audio. While there are no direct DG method comparisons available in this domain, our findings indicate that DG configurations can rival the effectiveness of several established UDA approaches that incorporate target data during training.

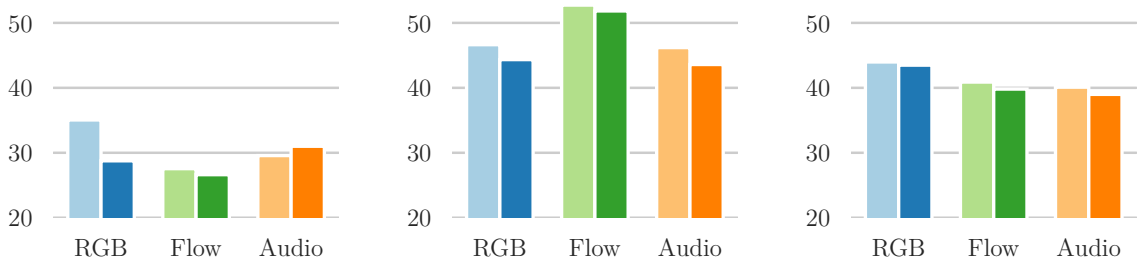
Method	Sampling	D1→D2	D1→D3	D2→D1	D2→D3	D3→D1	D3→D2	Mean
RGB + Flow								
Source Only	D-D	42.00	41.20	42.50	46.50	44.30	56.30	45.47
GRL (Ganin et al., 2016)	D-D	50.20	44.70	46.90	50.80	50.20	53.60	49.40
MMD (Long et al., 2015)	D-D	46.60	39.20	43.10	48.50	48.30	55.20	46.82
AdaBN (Li et al., 2018d)	D-D	47.00	40.30	44.60	48.80	47.80	54.70	47.20
MCD (Saito et al., 2018)	D-D	46.50	43.50	42.10	51.00	47.90	52.70	47.28
DAAA (Jamal et al., 2018)	D-D	50.00	43.50	46.50	51.50	51.00	53.70	49.37
MM-SADA (Munro and Damen, 2020a)	D-D	49.50	44.10	48.20	52.70	50.90	56.10	50.25
Kim et al. (Kim et al., 2021b)	D-D	50.30	46.30	49.50	52.00	51.50	56.30	50.98
STCDA (Song et al., 2021b)	D-D	52.00	45.50	49.00	52.50	52.60	55.60	<u>51.20</u>
Our (UDA)	D-D	50.84	47.14	48.86	54.38	50.60	58.43	**51.71**
Source Only	U-U	43.20	42.50	43.00	48.00	43.00	55.50	45.90
CIA (Yang et al., 2022a)	U-U	52.50	47.80	49.80	53.20	52.20	57.60	<u>52.18</u>
Our (UDA)	U-U	52.84	47.49	54.41	54.11	55.53	61.64	**54.34**
Source Only	D-U	54.25	50.72	54.87	56.41	51.65	61.27	54.86
Our (DG)	D-U	56.00	50.39	56.25	56.37	56.73	61.63	<u>56.23</u>
Our (UDA)	D-U	57.33	52.84	57.19	56.78	57.27	62.03	**57.24**
RGB + Audio								
Source Only	D-D	39.03	39.17	35.27	47.52	40.25	49.98	41.87
GRL (Ganin et al., 2016)	D-D	41.02	43.04	39.36	49.25	38.77	50.56	43.67
MMD (Long et al., 2015)	D-D	42.40	43.84	40.87	48.13	41.46	50.03	44.46
AdaBN (Li et al., 2018d)	D-D	36.64	42.57	33.97	46.63	40.51	51.20	41.92
MM-SADA (Munro and Damen, 2020a)	D-D	48.90	46.66	39.51	50.89	45.42	55.14	<u>47.75</u>
Our (DG)	D-D	42.55	41.77	42.73	51.09	42.63	54.24	46.21
Our (UDA)	D-D	46.65	47.22	46.18	52.30	44.04	56.18	**48.76**

Table 3.2 **Results on EPIC-KITCHENS-55.** Classification accuracies (%) on EPIC-KITCHENS-55 (Damen et al., 2018), using the evaluation protocol from (Munro and Damen, 2020a), divided by modalities. Results are grouped by the sampling strategy used for a fair comparison. Best in **bold**, runner-up <u>underlined</u>.

3.3.3 Ablation studies

In this section, we discuss the ablation studies conducted for our approach, all of which were carried out using the EK100 dataset. Given that EK100 is the largest and most diverse benchmark used in our work, it enhances the statistical significance of these studies.

Global alignment: a qualitative analysis. In Figure 3.6, we present the average feature norms for each modality. For simplicity, our discussion will focus on the verb feature norms, as the same observations are applicable to noun features. Specifically, Figure 3.6 illustrates how the average norms of verb features across different modalities vary within Domain Generalization (DG) and Unsupervised Domain Adaptation (UDA) scenarios, highlighting the influence of $\mathcal{L}_{RNA}$.



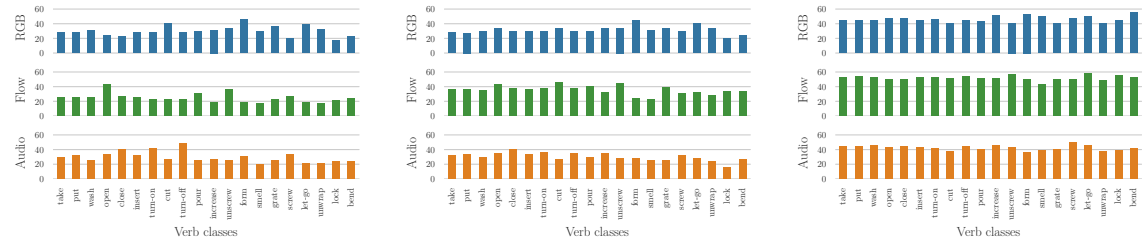
(a) Source only (b) $\mathcal{L}_{RNA}$ (DG) (c) $\mathcal{L}_{RNA}$ (UDA)

Fig. 3.6 **Verb feature norms across different modalities and settings (DG and UDA).** Light ( ) and dark colors ( ) indicate source and target validation domains, respectively. **(a)** In the Source Only configuration, distinct modalities and domains exhibit imbalanced feature norms. **(b)** $\mathcal{L}_{RNA}$ in DG enhances the alignment between different modalities, but a discrepancy between the source and target domains still remains. **(c)** Finally, the inclusion of $\mathcal{L}^{mod}$ in $\mathcal{L}_{RNA}$ reduces this gap in UDA, resulting in more uniform feature norms across different modalities and domains.

A preliminary qualitative assessment of the data depicted in Figure 3.6 indicates that $\mathcal{L}_{RNA}$ within the DG context (as shown in Figure3.6-b) results in improved alignment of the average feature norms across different modalities and an overall elevation in their values compared to the Source Only scenario (depicted in Figure 3.6-a). It is important to note that the norm formulation in Eq. 3.6 aims to address the alignment challenge at the batch level, therefore it does not assure a precise alignment of all average norms. Additionally, Figure 3.6-b reveals an increase in the Flow norm within DG in comparison to the Source Only condition (Figure 3.6-a). Prior research has demonstrated that the Flow modality is least impacted by domain shifts in egocentric action recognition (Munro and Damen, 2020a), which could potentially enhance generalization capabilities. This may account for the network's increased focus on this modality in the DG setting.

In addition, the presence of target data in UDA allows $\mathcal{L}_{RNA}$ to enhance the equilibrium among the norms of the different modalities, facilitating the model's ability to optimally leverage each modality's contributions for its final decisions. The improved complementarity between modalities, as evidenced by the increased accuracy shown in Table 3.3, may explain the (relatively) lower norm of Flow in UDA. This is counterbalanced by heightened norms for (and thus, increased emphasis on) the other two modalities, RGB and Audio.

Class alignment. To evaluate the impact of $\mathcal{L}^{c}_{RNA}$, we illustrate in Figure 3.7 the evolution of verb norms for both the ten most frequent and the least frequent classes in the DG context. In the Source Only scenario (Figure 3.7-a), the mean norms of features per class are



(a) Source Only (b) $\mathcal{L}^{g}_{RNA}$ (c) $\mathcal{L}^{g}_{RNA} + \mathcal{L}^{c}_{RNA}$

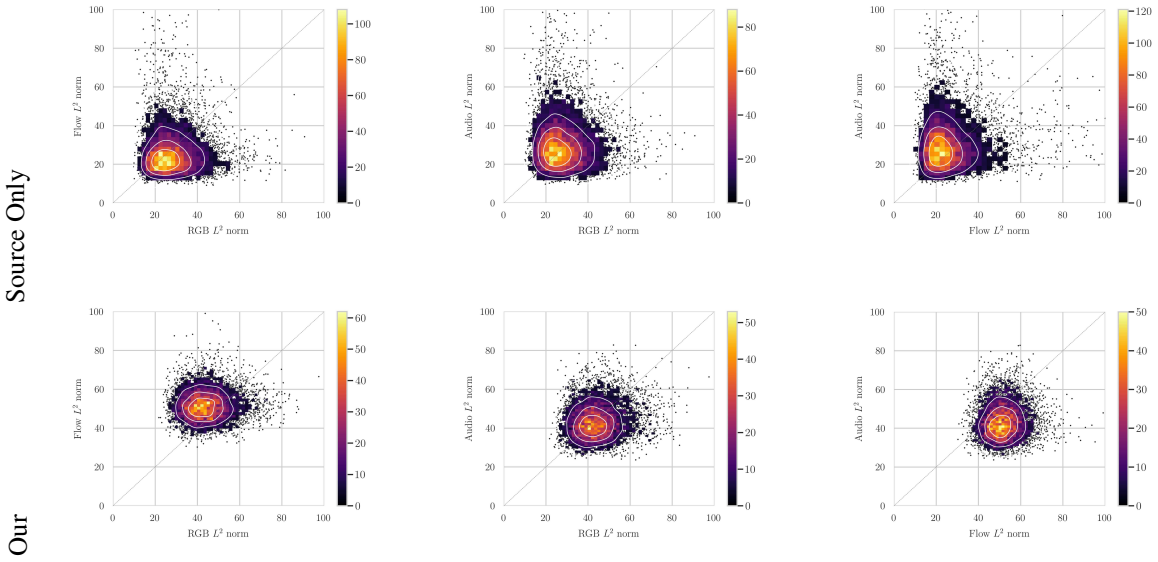
Fig. 3.7 **Per-class feature norms.** Feature norms for the top 10 most and least common classes from the target validation split of EPIC-Kitchens-100 are examined. Although $\mathcal{L}^{g}_{RNA}$ enhances the alignment across various modalities, a discrepancy among classes remains evident. Integrating the per-class variant of RNA significantly improves this misalignment, leading to more uniform feature norms across diverse classes.

notably imbalanced. Although the exclusive application of $\mathcal{L}^{g}_{RNA}$ achieves a more uniform balance across modality norms, its influence on equalizing the norms on a per-class basis remains minimal (Figure3.7-b). Conversely, the addition of $\mathcal{L}^{c}_{RNA}$ to the minimization process markedly enhances their alignment (Figure3.7-c), showcasing a substantial improvement in the balance of per-class feature norms.

Overall effect on feature norms. To delve deeper into the effects of $\mathcal{L}_{RNA}$, Figure 3.8 presents a scatter plot for the validation set under the DG setting. This visual representation is obtained by plotting the feature norms for RGB, Flow, and Audio of each sample within a three-dimensional space, where the axes correspond to the norms for the three modalities. To simplify interpretation, rather than offering a singular 3D plot, the data is shown through three distinct projections along the coordinate planes formed by the modality pairs. The objective of these visualizations is to illustrate the alterations in the manifold's configuration resulting from the application of $\mathcal{L}_{RNA}$.

The Source Only features display a wide dispersion, reflecting a manifold with a largely irregular configuration. This irregularity is attributed to the lack of alignment among the feature norms across different modalities. Using $\mathcal{L}_{RNA}$, the manifold has a more spherical and compact form, signifying improved alignment of modality norms. Moreover, there is a noticeable increase in the average feature norm values, causing the manifold to shift towards the upper right quadrant in the 2D visualizations.

Effect of loss components. Table 3.3 outlines how various loss components contribute to the final performance in both DG and UDA scenarios. To easily highlight the effect



(a) Flow v. RGB (b) Audio v. RGB (c) Audio v. Flow

Fig. 3.8 **Comparison of the feature norms before (top) and after (bottom) application of $\mathcal{L}^g_{RNA}$ and $\mathcal{L}^c_{RNA}$.** Each dot in the plots represents a sample from the validation dataset, with the color bar indicating increasing density values. Initially, the Source Only features exhibit a broad spectrum of values and an irregular configuration, highlighting the disparity in feature norms across the modalities. The introduction of the RNA loss readjusts this balance, leading to a more spherical distribution and concurrently increases the average norms.

of each component, we report the average improvement in terms of accuracy across verb, noun and action metrics (Δ Acc.). Integrating both global and class components in the DG setting yields a notable increase in accuracy (+2.20%) compared to using $\mathcal{L}^g_{RNA}$ alone (+1.36%). This demonstrates that combining these two components is effective in mitigating domain shift. Furthermore, the utilization of target data in UDA enhances the accuracy improvement to 1.78% for $\mathcal{L}^g_{RNA}$ and to 2.28% for $\mathcal{L}^g_{RNA}$ + $\mathcal{L}^c_{RNA}$, with the addition of $\mathcal{L}^{mod}_{RNA}$ further elevating the average improvement to 2.48%.

As detailed in Section 3.2.2, the UDA learning objective is enhanced by two additional loss functions: the adversarial domain loss $\mathcal{L}_d$, which seeks to improve feature transferability across domains, and the Information Maximisation loss $\mathcal{L}_{IM}$, aimed at reducing classification uncertainty among target classes. In this specific context, $\mathcal{L}_d$ leads to a more substantial improvement (2.71%), while $\mathcal{L}_{IM}$ has a lesser impact on overall accuracy. However, it is important to highlight that the combined effect of these two terms ($\mathcal{L}_d$ and $\mathcal{L}_{IM}$) varies depending on the specific task and benchmark, with some experiments demonstrating more significant benefits from $\mathcal{L}_{IM}$.

Method	Verb@1	Noun@1	Action@1	Δ Acc.
Source only	46.79	26.79	18.29	-
DG				
$\mathcal{L}^g_{RNA}$	49.53	27.50	18.91	1.36
$\mathcal{L}^g_{RNA} + \mathcal{L}^c_{RNA}$	<u>50.75</u>	27.92	19.81	2.20
UDA				
$\mathcal{L}^g_{RNA}$	49.98	27.79	19.44	1.78
$\mathcal{L}^g_{RNA} + \mathcal{L}^c_{RNA}$	50.46	28.49	19.77	2.28
$\mathcal{L}^g_{RNA} + \mathcal{L}^c_{RNA} + \mathcal{L}^{mod}_{RNA}$	49.94	**29.48**	19.87	2.48
$\mathcal{L}_{RNA} + \mathcal{L}_d$	50.59	<u>29.38</u>	<u>20.04</u>	2.71
$\mathcal{L}_{RNA} + \mathcal{L}_d + \mathcal{L}_{IM}$	**50.82**	29.19	**20.05**	2.73

Table 3.3 **Ablation on different loss components.** Δ Acc. is the average accuracy improvement for the verb, noun, and action metrics. Best in **bold** and the runner-up <u>underlined</u>.

Multi-modal adaptation capabilities. Another interesting question is whether the proposed method facilitates effective integration of multiple modalities in the final decision-making process, and if leveraging multiple modalities also enhances the domain adaptation capabilities of the model.

Table 3.4 presents a comparison of results obtained from experiments using pairs of modalities versus all three modalities together. It reveals that using all three modalities not only surpasses the performance of any pair of modalities but also demonstrates superior generalization capabilities. This is evidenced by an improved delta compared to the Source Only scenario (2.73%) versus the best two-modality improvement, achieved with Flow + Audio (2.06%). These findings indicate that our method is successful in combining different modalities to boost both the overall accuracy and the generalizability of the derived features.

Modality drop. In Table 3.5, we present an experiment designed to explore the impact of modality imbalance during training. Specifically, we examine the scenario in which a modality is "unexpectedly" lost at inference time, without the training strategy being designed to accommodate such a possibility. This situation, also discussed in (Gong et al., 2023), is significant because constraints at inference time—such as power, computational, or privacy constraints, or anomalies in an input device—might prevent the use of all modalities.

The core concept of our method is to enable the model to learn from different modalities equitably by re-balancing their contributions. Although it is evident that an unexpected loss of a modality results in decreased accuracy, we hypothesize that the influence of RNA

Method	Verb@1	Noun@1	Action@1	Δ Acc.
RGB + Flow				
Source Only	44.80	25.35	16.33	-
Our (DG)	45.95	26.65	16.94	1.02
Our (UDA)	47.64	26.49	16.91	1.52
RGB + Audio				
Source Only	39.91	24.18	14.84	-
Our (DG)	42.04	25.54	15.67	1.44
Our (UDA)	42.26	26.45	15.98	1.92
Flow + Audio				
Source Only	45.11	21.98	15.37	-
Our (DG)	48.87	23.44	16.49	2.12
Our (UDA)	48.42	23.51	16.71	2.06
RGB + Flow + Audio				
Source Only	46.79	26.79	18.29	-
Our (DG)	50.75	27.92	19.81	2.20
Our (UDA)	50.82	29.19	20.05	2.73

Table 3.4 **Modality ablation.** Top-1 classification accuracies (%) on modality pairs on EPIC-Kitchens-100 (Damen et al., 2022). Δ Acc. is the average accuracy improvement for the verb, noun and action metrics.

Method	Verb@1	Noun@1	Action@1	Δ Acc.
No Audio @ Test				
Source only	41.61	21.91	13.07	-
DG	44.03	24.44	14.89	2.26
UDA ($\mathcal{L}_{RNA}$)	44.08	24.77	15.25	2.50
No Flow @ Test				
Source only	30.58	20.33	10.63	-
DG	36.88	22.82	12.89	3.69
UDA ($\mathcal{L}_{RNA}$)	36.67	21.83	12.46	3.14
No RGB @ Test				
Source only	37.69	17.99	12.41	-
DG	46.70	18.92	13.53	3.69
UDA ($\mathcal{L}_{RNA}$)	46.51	19.37	13.55	3.78

Table 3.5 **Modality drop.** All configurations are trained on all input modalities. At inference time, we simulate the loss of a modality, resulting in large performance drops that RNA helps mitigate.

serves to enhance the model's resilience to such modality drops more effectively than the Source Only model. The latter is less able to exploit the synergies between modalities and, consequently, is more susceptible to the influence of dominant modalities. This hypothesis is supported by the findings in Table 3.5, aligning with the insights from (Gong et al., 2023). These findings illustrate distinct yet consistent impacts on Source Only when various modalities are omitted at test time, notably significant declines in accuracy compared to the data in Table 3.3. Moreover, these results indicate that RNA's balancing effect may assist the model in mitigating the adverse effects of a missing modality by optimizing the combined contribution of the remaining modalities.

3.4 Conclusion

This chapter presents a strategy for tackling the challenge of multi-modal domain generalization and adaptation. Our approach is inspired by the observation that discrepancies in the marginal distributions of modalities can profoundly impact the training process, resulting in sub-optimal performance and disparities in feature norms. To address these issues, we introduce the Relative Norm Alignment (RNA) loss, designed to re-balance the norms of features extracted across different domains and modalities, thereby enhancing overall accuracy. In UDA scenarios, this loss is synergized with adversarial domain loss and Information Maximization to boost feature transferability and regularization in the target domain. Our empirical findings demonstrate that the RNA method either surpasses or is on par with various state-of-the-art methods across egocentric action classification tasks, confirming its efficacy and versatility. Our method stands out for its simplicity and minimalistic design, facilitating easy integration into diverse architectures and settings without necessitating intricate adjustments. This inherent flexibility positions RNA as a viable option for real-world scenarios characterized by multi-modal data. Future research will delve into expanding RNA's utility and adaptability across a broader spectrum of domains and modalities. It will focus on overcoming issues related to imbalanced data distributions and will explore potential synergies with other techniques aimed at mitigating domain shifts and enhancing generalization.

A limitation we observed stems from the fact that in many real-world scenarios, data distributions are significantly imbalanced, which leads to reduced accuracy for the tail classes (Buda et al., 2018). Research illustrates how this imbalance results in uneven norms of classification weights per class (Guo and Zhang, 2017; Kim and Kim, 2020), as well as imbalanced norms of features per class (Li et al., 2022a; Wu et al., 2017). In developing our

method, we hypothesized that equalizing the norms per class would also positively affect the re-balancing of the classifier's weights for the tail classes. However, our experimental findings reveal that this anticipated effect does not occur. This revelation opens avenues for future research to integrate this goal into RNA as an additional component for re-balancing the classifier's weights.

Chapter 4

Vision and Language for Domain Generalization

The previous chapter focuses on a Domain Generalization (DG) setting, where the domain gap between the source and target domains is mainly due to differences in the environments where the activities occur, though still limited to "cooking" activities. In fact, until now, research efforts in DG have predominantly addressed generalization across visual domain shifts (Damen et al., 2018; Li et al., 2017a; Munro and Damen, 2020a; Planamente et al., 2022b; Torralba and Efros, 2011). These studies have sought to understand how models can be adapted to perform accurately across diverse visual environments that they were not specifically trained on. While valuable, this approach to DG has mainly focused on variations in appearance, lighting, or background across datasets.

In this chapter, we delve into the concept of *scenario shift*, a relatively unexplored dimension of DG. Scenario shift involves situations where the same action — such as cutting, moving, or assembling — is carried out in completely different contexts or activities. This introduces variations not just in the visual domain but in the functional context of the action, including changes in the tools used, the objects being interacted with, and even the ultimate goals and expected outcomes of these actions. We also investigate the impact of *location shift*, another critical but underexplored factor in DG. Location shift recognizes that the same action can be executed differently across various geographic and cultural contexts, influenced by local customs, available materials, and environmental conditions.

To facilitate our analysis, we introduce a specialized dataset named ARGO1M. This collection features 1.1 million action clips spanning 60 different classes, sourced from 73 unique scenario/location combinations.

By combining video data with its associated narrations, we can leverage the complementary nature of visual and language information. The visual content offers insights into the physical actions and interactions, while the textual descriptions enrich this understanding by providing a semantic layer, clarifying motivations, contexts, and relationships. This dual-source approach facilitates a more comprehensive understanding of the video's content, improving the model's ability to generalize across different domains and contexts.

The work presented in this chapter led to the publication of one work:

- Plizzari, C., Perrett, T., Caputo, B., & Damen, D. (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13656-13666)*.
  Online Resources: [Paper], [Project page]

4.1 Introduction

A notable distinction between human intelligence and artificial intelligence lies in the remarkable human capacity for generalization. Consider observing the action of "cutting" as executed by a chef in Italy; we can instinctively recognize the same action when it is performed in a geographically distinct *location*, such as India, despite never having physically visited the location. Furthermore, our cognitive abilities allow us to identify actions across novel *scenarios*. For instance, we can understand a mechanic cutting metal, even if we have no prior experience with the tools they are using.

This challenge is encapsulated in the concept of domain generalization (Zhou et al., 2022), which describes a scenario where a model, trained on a specific set of labeled data, struggles to apply its learning to a different, unseen data distribution during inference. The gap between these distributions is termed as *domain shift*. Research in domain generalization has largely focused on navigating visual domain shifts (Damen et al., 2018; Li et al., 2017a; Munro and Damen, 2020a; Planamente et al., 2022b; Torralba and Efros, 2011).

This work delves into the concept of *scenario shift*, wherein the same action occurs within different activities, thus influencing the tools used, the objects interacted with, and the ultimate goals and behaviors. For instance, the act of "cutting" could involve various tools (such as scissors, knives, or saws) and objects (ranging from paper and vegetables to wood), depending on the context, whether it be in a kitchen, a workshop, or an art studio, each presenting unique challenges for recognizing the action.

Moreover, we explore the notion of *location shift*, acknowledging that identical actions may be performed differently across diverse geographic and cultural contexts, shaped by local traditions, available resources, and environmental conditions.

In Figure 4.1, we illustrate the action "cut" being performed with a knife in cooking (), with pliers in construction (), and with scissors in arts and crafts (). The choice of tools is not restricted to a particular scenario and can vary between locations — for instance, as depicted in Figure 4.1, scissors are used to cut seaweed sheets while cooking in Japan. Optimal generalization would involve understanding the essence of "cutting" as the act of dividing an object into two or more sections, independent of the tool used or the setting. Such generalization capabilities could facilitate the recognition of metal being "cut" by a mechanic in India using an angle grinder (Figure 4.1, Test), showcasing successful domain generalization.

Fig. 4.1 **Problem statement.** Problem statement and examples from the ARGO1M dataset illustrate that the same action, e.g., "cut", can be executed differently depending on the *scenario* and *location* where it takes place. Our objective is to generalize such that we can recognize the same action within a new scenario, *unseen* during training, and in an *unseen location*, for instance, a *Mechanic* (🔨) in *India* (🇮🇳).

Our research is made possible by the recent release of the Ego4D dataset (Grauman et al., 2022), which provides egocentric footage from across the world. We have created a specific dataset for action generalization, named ARGO1M. This dataset comprises 1.1 million action clips across 60 classes, originating from 73 unique scenario/location pairings.

To address the challenges presented by ARGO1M, we introduce a novel approach for domain generalization. Our method models each video as a weighted combination of other videos within the same batch, which may belong to different domains. This technique is termed Cross-Instance Reconstruction (CIR). Through the process of reconstruction, CIR learns to extract video features that can be generalized across various domains. The supervision of CIR involves both a classification loss and a video-text association loss, enabling it to effectively learn domain-invariant features through language. The classification loss guides the model to accurately predict classes of actions within the video content, while the video-text association loss strengthens its ability to link visual content to corresponding textual descriptions. This dual strategy helps the model capture the relationships between video and language, thereby improving its ability to generalize across different domains.

To summarize, the contributions of this chapter are:

- We curate the Action Recognition Generalization dataset (ARGO1M) utilizing videos and narrations from Ego4D (Grauman et al., 2022). This dataset, ARGO1M, stands as

the first dataset designed to evaluate action Generalization across both scenario and location shifts, making it the most extensive domain Generalization dataset for both images and videos to date (Section 4.3).

- We present CIR, a domain generalization approach that leverages Cross-Instance Reconstruction along with video-text pairing to learn generalizable representations (Section 4.4).
- We evaluated CIR on the proposed ARGO1M, demonstrating that it consistently surpasses both baseline models and recent domain generalization techniques across 10 test sets (Section 4.5).

4.2 Background

In this section, we review existing datasets for Domain Generalization (DG) and existing approaches performing cross-instance reconstruction tasks. Note that DG aims to generalize to any unseen target domain, without having access to data from that target domain during the training phase (Zhou et al., 2022). This is different from the Domain Adaptation approach, where unlabeled target domain samples are accessible during training (Kim et al., 2021a; Munro and Damen, 2020a; Song et al., 2021a). For a more comprehensive discussion on the distinctions between these two approaches, we direct readers to Section 4.3.

Domain Generalization (DG) datasets. Table 4.1 provides a detailed comparison of various vision datasets that have been curated for the purpose of domain generalization research. These existing image datasets predominantly feature a stylistic variation across their contents. Datasets such as PACS (Li et al., 2017a), Office-Home (Venkateswara et al., 2017), and DomainNet (Peng et al., 2019) include a diverse range of common objects depicted in different artistic styles including photos, paintings, clipart, cartoons, and sketches. This approach to dataset composition illustrates the exploration of stylistic shifts within the data, showcasing common objects and categories represented across a variety of artistic expressions (Li et al., 2017a; Peng et al., 2019; Venkateswara et al., 2017), as well as across different datasets (Torralba and Efros, 2011). The concept of location shift has been explored in (Beery et al., 2018), which contains images of animals captured in a variety of geographical settings.

In videos, domain shifts are characterized by various factors including cross-dataset variations (Chen et al., 2019), transitions from synthetic to real environments (Chen et al.,

		Samples		Domains		
	Dataset	# Samples	# Cls	# Train	# Test	Domain Shift
Images	PACS (Li et al., 2017a)	9,991	7	3	4	Style
	VLCS (Torralba and Efros, 2011)	10,729	5	3	4	N/A
	OfficeHome (Venkateswara et al., 2017)	15,588	65	3	4	Style
	TerraIncognita (Beery et al., 2018)	24,788	10	3	4	Loc
	DomainNet (Peng et al., 2019)	586,575	345	5	6	Style
Videos	UCF-HMDB (Chen et al., 2019)	3809	12	1	2	N/A
	Kinetics-Gameplay (Chen et al., 2019)	49,998	30	1	2	Realism
	MM-SADA (Munro and Damen, 2020a)	10,094	8	2	3	Loc
	EPIC-Kitchens (Damen et al., 2022)	48,139	86	11	1	Time Gap
	ARGO1M	1,050,371	60	54-64	10	(Scenario, Loc)

Table 4.1 **Datasets for DG.** ARGO1M offers combined scenario and location shifts, and is the largest DG dataset in terms of # of samples and # of domains.

2019), changes in viewpoint (Choi et al., 2020a), geographical location differences (Munro and Damen, 2020a), and even the effects of time progression (Damen et al., 2022).

ARGO1M is 21× larger than any existing video DG dataset and 1.8× larger than any image DG dataset previously reported. Critically, ARGO1M introduces the concept of scenario shift. This involves testing the generalization of models not just across different locations but also across a wide range of scenarios, featuring an unprecedented scale of domain diversity with up to 64 training domains and 10 test domains.

Cross-Attention for Reconstruction. The approach of predicting masked tokens within a video has become a widespread technique in many representation learning methods (Feichtenhofer et al., 2022). (Feichtenhofer et al., 2022) randomly mask out space-time patches in videos and train an autoencoder to reconstruct them. They show that this method can learn strong spatiotemporal representations from videos with almost no domain-specific bias. Our method differs from these strategies by focusing on reconstruction using other videos within the same batch. This idea of cross-instance attention, where query instances are reconstructed from examples of each class, has seen application in few-shot learning (Doersch et al., 2020; Perrett et al., 2021). In the work of (Perrett et al., 2023), instances for few-shot classes are reconstructed from samples belonging to head classes. This has been shown to improve performance on long-tail video recognition. Similarly, in cross-modal retrieval (Patrick et al., 2020), reconstruction through cross-attention aids in enhancing video-text representations via a caption generation task. Specifically, each video's caption is reconstructed as a weighted combination of other support videos' visual representations.

Unlike these previous approaches, our method is unique in that it reconstructs each video as a learned weighted mixture of videos from *various domains*. This introduces a novel

narration: 01:55 #C C trims the plant stem with the pruner in his right hand
location: unict
scenario: gardening

narration: 00:39 #C C stirs food
location: utokyo
scenario: cooking

narration: 27:17 #C C shuffle cards
location: bristol
scenario: playing board games

narration: 05:56 #C C unwraps the black tape with his hands
location: cmu
scenario: car mechanic

Fig. 4.2 **Samples from Ego4D.** Each video clip is associated to a timestamp and narration, the geographic location where the video was captured, and a scenario.

aspect to the domain generalization challenge, as it allows to learn robust representations that can better generalize to unseen domains.

4.3 ARGO1M Benchmark

In this section, we describe how we curated the ARGO1M dataset from videos within the Ego4D dataset (Grauman et al., 2022).

Ego4D Background. The Ego4D dataset (Grauman et al., 2022) comprises unedited egocentric videos amounting to 3,670 hours, recorded in eight countries outside the US and five US states. These videos encapsulate a diverse array of everyday life situations, such as playing cards, cooking, and car repair. Each video comes with metadata detailing its geographic location and the scenario depicted. The majority of the videos focus on a single scenario, although 14.9% are noted to include multiple scenarios. Narrations at the timestamp level are available within each video, outlining the actions and object interactions of the person wearing the camera. An example of such a narration is "#C C puts the scraper down" at the timestamp of 3.70 seconds. Some examples of video clips from Ego4D and the associated metadata are shown in Figure 4.2.

ARGO1M scenarios. The high-level scenario descriptions in Ego4D are often in free-form and sometimes absent. We exclude Ego4D videos lacking a scenario description (7.4% of the total videos). From the 136 free-form scenario descriptions provided by Ego4D, we select 62 that offer a significant diversity and volume of videos. We omit those that are redundant or not indicative of a distinct activity, such as "Talking" or "On a screen". This process yields a collection of 6,813 videos, accounting for 83.1% of the videos associated with at least one scenario. Videos identified as encompassing multiple scenarios are also excluded. Subsequently, we manually categorize the free-form descriptions into 10 distinct

Scenario	Ego4D Descriptions
Cooking	BBQing/picnics, Baker, Cooking, Making coffee, Outdoor cooking
Building	Carpenter, Fixing something in the home, Handyman, Making bricks, Jobs related to construction/renovation company (director of work, tiler, plumber, electrician, handyman, etc)
Arts and crafts	Crafting/knitting/sewing/drawing/painting
Cleaning	Car/scooter washing, Cleaning / laundry, Cleaning at the gym, Community cleaning, Daily hygiene, Household cleaners, Washing the dog / pet or grooming horse
Mechanic	Assembling furniture, Bike mechanic, Blacksmith, Car mechanic, Fixing PC, Getting car fixed, Labwork, Maker Lab (making items in different materials, wood plastic and also electronics)- some overlap with construction etc. but benefit is all activities take place within a few rooms, Scooter mechanic, Working at desk, Biology experiments
Gardening	Doing yardwork / shoveling snow, Farmer, Flower picking, Gardener, Gardening, Potting plants (indoor)
Playing	Assembling a puzzle, Gaming arcade / pool / billiards, Playing darts, Playing board games, Playing cards, Playing games / video games, Practicing a musical instrument
Shopping	Clothes and other shopping, Grocery shopping indoors, Working in milktea shop, Working in outdoor store
Sport	Attending sporting events - watching and participating in, Baseball, Basketball, Bowling, Climbing, Cycling / jogging, Football, Going to the gym - (exercise machine, class, weights), Golfing, Hiking, Playing badminton, Roller skating, Rowing, Swimming in a pool/ocean, Working out at home, Working out outside
Knitting	All videos from *Arts and crafts* scenario, where *at least* one narration contains keywords related to knitting activities.

Table 4.2 **Closed-form scenarios for ARGO1M**, and corresponding Ego4D free-form descriptions.

scenarios: *Cooking* (), *Building* (), *Arts and Crafts* (), *Cleaning* (), *Mechanic* (), *Gardening* (), *Playing* (), *Shopping* (), *Sport* (), and *Knitting* (). For instance, the descriptions "Car mechanic", "Getting the car fixed", and "Bike mechanic" are grouped under *Mechanic*. The clustered scenarios are detailed in Table 4.2.

ARGO1M video clips. Each chosen video comes with detailed timestamp-level narrations that describes the actions and object interactions of the person wearing the camera. For instance, the narration "#C C puts the scraper down" is noted at the 3.70s timestamp. We selected narrations from *annotator_1*, focusing exclusively on actions performed by the camera wearer, indicated by narrations tagged with `#C`, while disregarding actions by external actors, which are marked with `#O`. Additionally, we employed a series of heuristics to filter out videos with inaccurately provided scenario metadata by Ego4D. This process involved pinpointing a set of keywords expected to be present in narrations corresponding to the scenario across the videos. We kept videos whose narrations included these scenario-specific keywords that we manually identified, resulting in a refined collection of 6,358 videos (93% of the videos from the designated scenarios) with 1,637,810 narrations. The narrations in Ego4D are accurately synchronized with the videos, thanks to a pause-and-narrate annotation

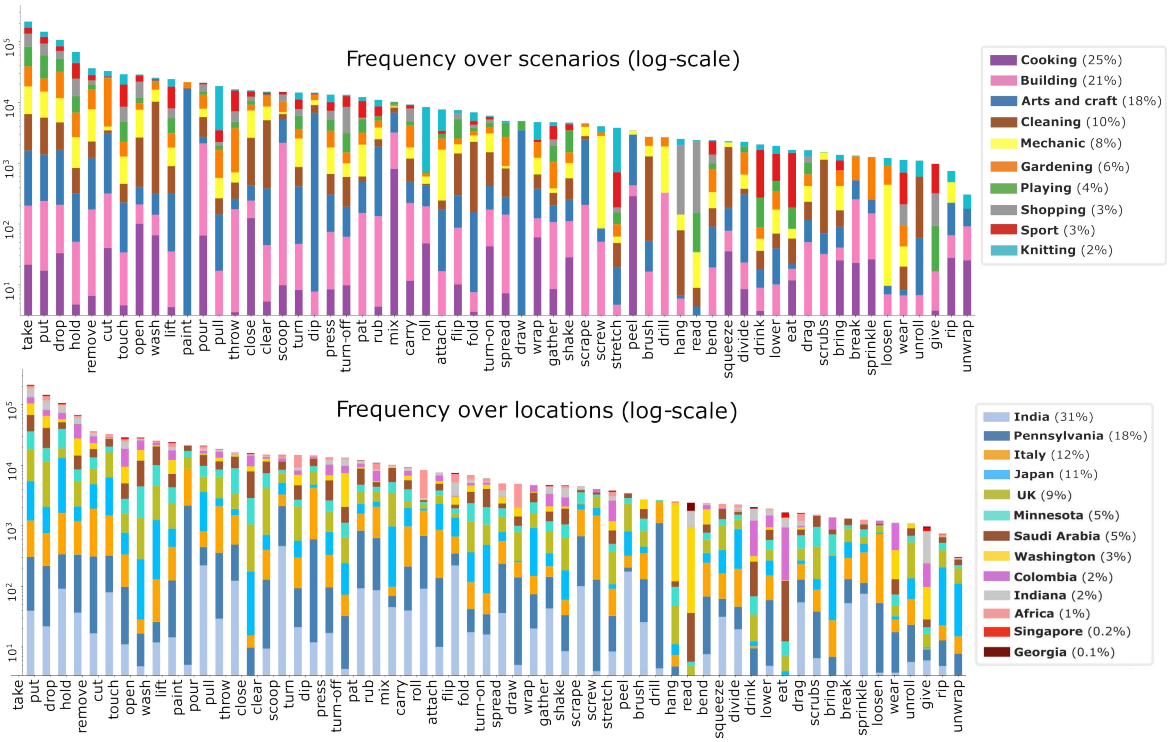


Fig. 4.3 **Per-class distribution.** The frequency (on a log scale) of the 60 classes within ARGO1M is depicted across scenarios (top) and locations (bottom), with percentages indicated in the legend. Within each bar, both scenarios and locations are linearly.

method, as reported in the Ego4D paper and by other studies (Lin et al., 2022). To confirm this, we manually annotated the start times of actions in a small sample and observed an average discrepancy of 0.6s between our noted action start times and the provided narration timestamps, and a 0.9s difference for their conclusions. This precision enables us to consider the narration timestamp as the beginning of the clip, and the timestamp of the subsequent narration as the clip's endpoint. Following previous works, where action boundaries may be loosely defined as long as they encapsulate the pertinent action (for example, as seen in Kinetics (Carreira and Zisserman, 2017)), we consider these boundaries adequate for both training and evaluating action recognition models.We next describe how clips are associated with class labels.

ARGO1M action classes. Action classes are obtained from the verbs in Ego4D narrations using the spaCy (Honnibal and Montani, 2017) tool. Narrations are parsed to identify verbs and nouns, with verbs considered as potential actions. These verbs are then categorised according to the taxonomy from EPIC-KITCHENS-100 (Damen et al., 2022), albeit with some modifications to accommodate the broader spectrum of activities captured in Ego4D. For instance, similar actions such as "take" and "pick" are consolidated into a single class.

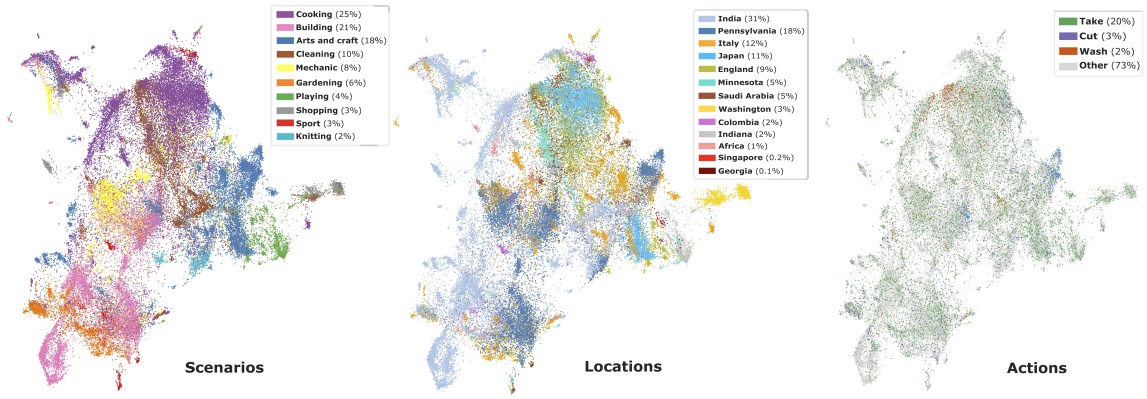


Fig. 4.4 **ARGO1M feature distribution.** UMAPs (Uniform Manifold Approximation and Projection) for ARGO1M features showcase the distribution across scenarios (left), locations (center), and for three specific action classes (right). To demonstrate the alignment across these three dimensions, the same projection is utilized across all three UMAP plots.

Actions that are ambiguous, *e.g.* "adjust", or that do not involve interaction with the environment, *e.g.* "look at", are omitted. Additionally, actions that occur too infrequently for effective domain generalization training are excluded. This curation results in a final set of 60 action classes (shown in Figure 4.3) including 1,050,371 instances. The distribution exhibits a long-tailed pattern, with each action class occurring across a variety of scenarios and locations, as illustrated in Figure 4.3. On average, each class is represented in 8 different scenarios and 11 distinct locations. Thus, ARGO1M comprises 1,050,371 video clips sourced from 5,894 videos, representing 42% of all clips within Ego4D and 61% of all videos selected from the dataset for this study.

In summary, ARGO1M contains 1,050,371 video clips. Each video *clip* is captured in a given *scenario* (out of 10) and geographic *location* (out of 13), with associated *text narration* and *action class* (out of 60). For example, the caption, "#Camera wearer (C) cuts the lemon strand." is associated to a clip recorded in "Italy" and capturing "Gardening" scenario, with associated action label "cut".

ARGO1M feature distribution. Figure 4.4 offers an insight into the feature distribution of all samples within ARGO1M, highlighting variations across scenarios (left), geographic locations (center), and action classes (right). For better visual clarity, in the action class diagram, we selectively depict 3 out of the 60 classes and categorize the rest under *others*. These features, derived from a SlowFast network (Feichtenhofer et al., 2019) that was pre-trained on Kinetics (Carreira and Zisserman, 2017), are visualized using UMAP.

There is noticeable evidence of *scenarios clustering according to different locations*, for example, the *Playing* scenario (indicated by a green cluster on the right side of the feature

map) spans various locations (*United Kingdom*, *Minnesota*, and *Indiana*), and *locations clustering around different scenarios*, such as *Minnesota* (highlighted by a yellow cluster on the right), which includes multiple scenarios, predominantly *Cleaning* and *Shopping*. This observation suggests that scenario and location shifts cannot be handled independently or disentangled easily. Therefore, considering (scenario, location) pairings as distinct test domains more accurately reflects the intricacies of combined scenario/location shifts.

Although clusters based on scenarios and locations are distinguishable, action classes appear more dispersed across the map. This dispersion is exemplified by the actions "take", "cut", and "wash", which are all widely scattered across the feature space. This dispersion underscores the complexity inherent in the generalization task at hand.

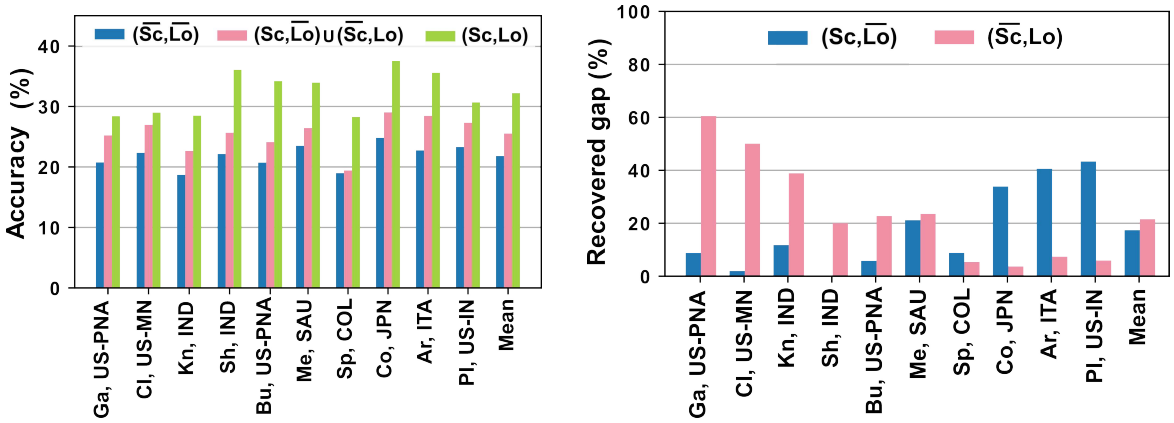
ARGO1M Splits. We curated 10 distinct train/test splits to assess generalization across both scenarios and locations. These 10 test splits are manually chosen to ensure coverage of *all scenarios*. For each scenario, we identify the location with the highest number of samples to create a test split that allows for a robust evaluation.

With a given pair of scenario and location **(Sc, Lo)**, the corresponding training split is designed to exclude all samples related to the scenario **(Sc)** as well as those from the location **(Lo)**. This methodology ensures a thorough generalization challenge by excluding examples from that scenario and location from the training set. Later in this section, we demonstrate that these 10 splits showcase a variety of combined scenario/location shift characteristics, highlighting the complexity of the generalization task at hand.

The selected test splits and their [number of samples] are: *Gardening* in *Pennsylvania* (**Ga, US-PNA**[1]) [16,410], *Cleaning* in *Minnesota* (**Cl, US-MN**) [22,008], *Knitting* in *India* (**Kn, IND**) [13,250], *Shopping* in *India* (**Sh, IND**) [11,239], *Building* in *Pennsylvania* (**Bu, US-PNA**) [99,865], *Mechanic* in *Saudi Arabia* (**Me, SAU**) [11,700], *Sport* in *Colombia* (**Sp, COL**) [16,453], *Cooking* in *Japan* (**Co, JPN**) [82,128], *Arts and crafts* in *Italy* (**Ar, ITA**) [36,812], *Playing* in *Indiana* (**Pl, US-IN**) [17,379].

ARGO1M Domain Shift analysis. We analyze the effects of scenario and location shifts within the 10 test splits of ARGO1M by varying the inclusion of samples from the test scenario and/or location in the training set. Throughout these experiments, we utilize Empirical Risk Minimization (ERM), which is standard cross-entropy training (refer to Section 4.5 for comprehensive experimental details). This preliminary analysis aims to shed light on the domain shift present in ARGO1M.

[1]We use ISO country codes and US state codes.



(a) Accuracy without samples from the test scenario or location **(Sc̄, Lō)** as well as **(Sc, Lō)∪(Sc̄, Lo)** and **(Sc Lo)**.

(b) % of drop recovered when adding examples from either scenario **(Sc, Lō)** or location **(Sc̄, Lo)**.

Fig. 4.5 **ARGO1M domain shifts.** Analysis of scenario and location shifts on ARGO1M.

Initially, we adopt the default setting (1), where samples from neither the test scenario nor the location are included in the training set, denoted as **(Sc̄, Lō)**, with the overline indicating exclusion from the training split. This is different from cases where (2) the training split incorporates samples featuring either the test scenario or location, but not both, indicated as **(Sc, Lō)∪(Sc̄, Lo)**, and (3) samples from both the test scenario and location are present, indicated as **(Sc, Lo)**. As illustrated in Figure 4.5a, performance improves from (1) → (2), with a more significant improvement observed from (2) → (3). This progression underscores the challenges in generalization when neither the test scenario nor location is represented during training.

Next, we examine the individual contributions of scenario and location shifts to the observed performance degradation. We explore the extent of recovery against (3) when incorporating training samples from either the test scenario **(Sc, Lō)** or location **(Sc̄, Lo)**. Figure 4.5b shows that the influence of scenario and location shifts varies significantly across the test splits. For instance, in the (**Sh, IND**) split, including the test scenario *shopping* yields no improvement, while incorporating the location *India* proves beneficial. In contrast, for (**Ar, ITA**), integrating *arts and crafts* mitigates 40% of the performance decline, whereas the location offers no advantage. This variation indicates the distinct challenges posed by both scenario and location shifts, with our 10 test splits providing a variety of cases to investigate these dynamics.

4.4 CIR: Cross-Instance Reconstruction

We introduce Cross-Instance Reconstruction (CIR) as a technique to represent an action by a weighted combination of actions from diverse scenarios and locations. In this section, we outline the inputs to our method and detail its specifics. We then describe the training process and the inference strategy used.

Proposed Setting. Each training instance consists of a video clip v, accompanied by a free-form text narration t, and an action class label y, denoted as (v, t, y). For testing purposes, the only requirement is the input video clip, from which the action label is predicted. We use $\hat{y}$ to denote the predicted label.

To classify actions, we employ a composite function:

$$\hat{y} = h \circ f(v) \tag{4.1}$$

Here, f represents an encoder that extracts a video representation suitable for domain generalization, and h is an action classifier operating on that representation.

In addition to the the cross-entropy loss $\mathcal{L}_c$ applied to h, we also train f utilizing two types of losses: a cross-modal loss and an additional classification loss.

Cross-Instance Reconstruction. The core idea behind cross-instance reconstruction (CIR) is to encourage cross-domain representations of actions, with domains defined as scenarios and locations. Through this approach, the representations become domain-generalizable as they reconstruct the same action using samples from other domains.

We adopt a learn-to-reconstruct approach for any given video clip using other video clips from a randomly sampled batch, referred to as the support set S. We reconstruct all video clips in the batch jointly at feature level. Consequently, each video clip is included in the support set for every other clip within the same batch. Before delving into the training objectives, we first explain the reconstruction process.

We learn two projection heads, designated as the query and key heads, Q and K, consistent with standard cross-attention methods (Vaswani et al., 2017), and implement a layer norm L. The correlation between each pair of video clips, v_i and v_j, in the training batch is computed as:

$$c_{ij} = L(Q(f(v_i))) \cdot L(K(f(v_j))) \tag{4.2}$$

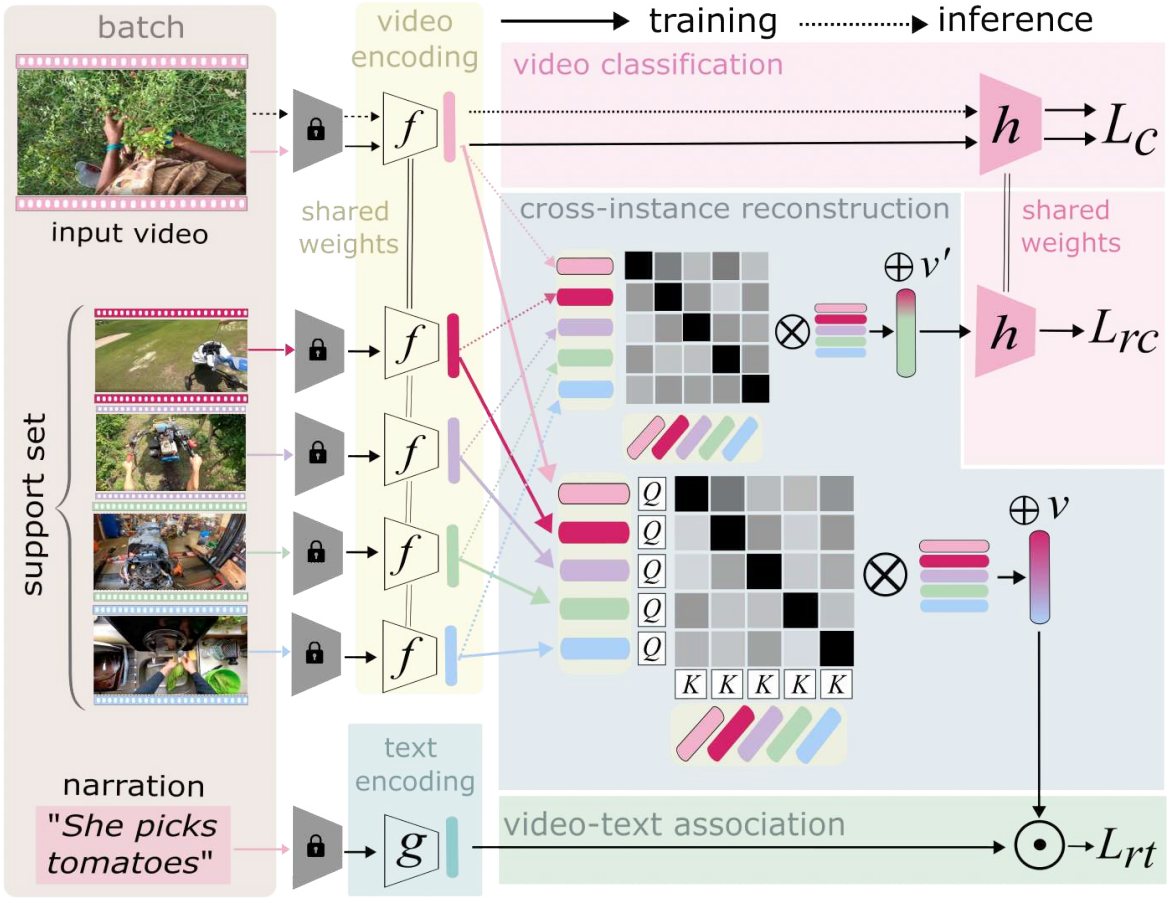


Fig. 4.6 **CIR**. A video clip and its corresponding narration are shown alongside the support set of other clips from the batch. Video $f(v)$ and text $g(t)$ embeddings are derived using trained encoders built upon a frozen model. The cross-entropy loss $\mathcal{L}_c$, along with two Cross-Instance Reconstruction (CIR) objectives $\mathcal{L}_{rt}$ and $\mathcal{L}_{rc}$, are minimized during training. For $\mathcal{L}_{rt}$, query Q and key K projections for clips within the batch are developed, with subsequent self-masking. The weights obtained are applied to $f(v)$, and the reconstructed $\oplus v$ is aligned with its corresponding narration. For $\mathcal{L}_{rc}$, the reconstructed $\oplus v'$ undergoes classification through the classifier h. During inference, only the video classifier h is utilized.

The computed weights c_{ij} undergo a softmax operation and are self-masked to prevent trivial reconstructions from the same sample. The reconstructed representation $\oplus v_i$ is the result of a weighted sum of all embeddings in its support set, based on the weights c_{ij}:

$$\forall i: \quad \oplus v_i = \sum_{j \in S} \frac{\exp(c_{ij}) f(v_j)}{\sum_{k \in S} \exp(c_{ik})} \tag{4.3}$$

We apply the weights directly to $f(v)$, which is analogous to using the identity matrix for the value head in traditional attention mechanisms.

Training CIR. Figure 4.6 provides an overview of CIR, which we next describe in more detail. Our goal is for the reconstructions to learn to generalize and then backpropagate this capability to the video encoder f (Eq. 4.1). We introduce two types of reconstructions, each driven by a distinct objective. The video-text association reconstruction ($\oplus v$ in Figure 4.6) leverages text narrations, thus enriching these cross-instance reconstructions with the semantic description of the video clip. Meanwhile, the classification reconstruction ($\oplus v'$ in Figure 4.6) is designed to identify the clip's action class. The former aims to reconstruct the specific instance of the action depicted in the video, whereas the latter focuses on cross-domain action level reconstructions.

For the **video-text association reconstruction** $\oplus v_i$, we employ contrastive learning to align $\oplus v_i$ closely with the text narration embedding associated with its video, for example, "He turns the lawn mower". Within a batch of video-text pairs and their corresponding reconstructions $\mathcal{B} = \{(v_i, \oplus v_i, t_i)\}_{i=1}^{B}$, we define the objective using Noise Contrastive Estimation (Oord et al., 2018) to focus on both reconstruction-text and text-reconstruction pairings. Particularly, the reconstruction-text loss considers the reconstruction $\oplus v_i$ as the anchor and the other text narrations in the batch as negatives, expressed as:

$$\mathcal{L}_{r\to t}(\oplus v_i, g(t_i)) = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp\left(s(\oplus v_i, g(t_i))/\tau\right)}{\sum_{j}^{B} \exp\Big(s(\oplus v_i, g(t_j))/\tau\Big)} \tag{4.4}$$

where $s(\cdot,\cdot)$ denotes cosine similarity, g represents the text encoder, $g(t_i)$ the encoded text narration, and τ a learnable temperature parameter. Conversely, the loss $\mathcal{L}_{t\to r}$ treats $g(t_i)$ as the anchor with other reconstructions acting as negatives. These components are illustrated in Figure 4.7. Together, they constitute our reconstruction-text association loss $\mathcal{L}_{rt} = \mathcal{L}_{r\to t} + \mathcal{L}_{t\to r}$.

Note that we avoid pairing this reconstruction with the video embedding $f(v_i)$, opting instead for the text narration $g(t_i)$. This is because the video embedding could convey domain knowledge (i.e., scenario and location), potentially biasing the reconstruction towards videos from the same scenario or location. Instead, the associated narration provides an instance-level description of the action, guiding the reconstruction more effectively.

Our **classification reconstruction** $\oplus v'_i$ serves as the input for the classifier h, enabling it to identify the action class such that $\hat{y}' = h(\oplus v')$. This process is guided by a cross-entropy loss, referred to as $\mathcal{L}_{rc}$, indicating its role in classifying reconstructions. The classifier's weights for videos and reconstructions are shared to maintain consistency.

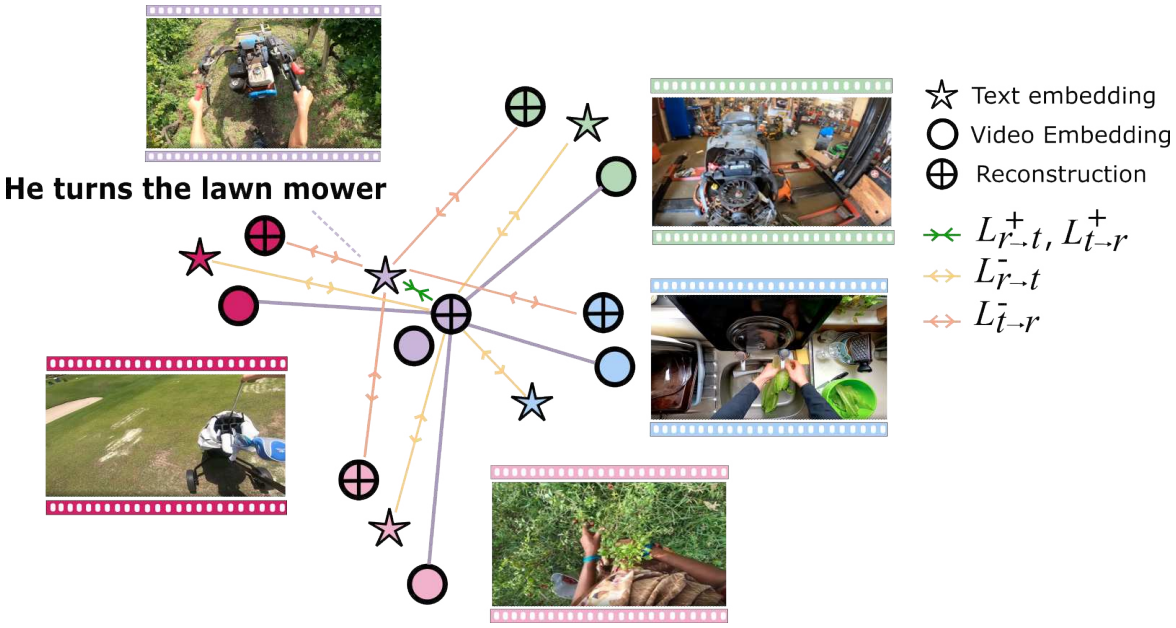


Fig. 4.7 **Video-text association.** The reconstructed clip $\oplus v'_i$ (violet) is matched with its text representation. The reconstruction-to-text loss $\mathcal{L}_{r\to t}$ treats $\oplus v'_i$ as the positive sample and other text narrations as negatives, while the text-to-reconstruction loss $\mathcal{L}_{t\to r}$ considers other reconstructions $\oplus v'_j$ as negatives.

Furthermore, for this particular reconstruction, we calculate weights using cross-product attention: $c'_{ij} = f(v_i) \cdot f(v_j)$, effectively replacing c with c' in Eq. 4.3. Consequently, we do not introduce separate query and key projections for this task. The rationale and impact of these choices are further examined in Section 4.5.2.

We integrate our two losses with the cross-entropy video classification loss $\mathcal{L}_c$ (detailed in Section 4.4) to form our comprehensive training objective:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_{rt} + \lambda_2 \mathcal{L}_{rc}. \tag{4.5}$$

Here, λ_1 and λ_2 are weights assigned to the two reconstruction losses, balancing their contribution to the overall training objective.

Inference. Once training is complete, f is capable of deriving domain-generalizable representations that encapsulate action class knowledge while remaining free of domain bias. Consequently, during the testing phase, only video clips v_i from the test split undergo processing by the encoder f and the classifier h. Narrations are not needed at this stage, and there is no reconstruction process involved – meaning each clip is classified independently.

4.5 Experiments

In this section, we evaluate the ability of CIR to generalize over scenarios and locations by comparing it against baseline and state-of-the-art domain generalization methods adapted for our setting. We then conduct ablation studies on its various components to understand their individual contributions. Additionally, we provide qualitative examples to visualize the effects of CIR.

Dataset and metrics. For performance evaluation, we conduct experiments on the 10 distinct test splits outlined in Section 4.3. We report the top-1 accuracy for each test split, as well as the average accuracy. We employ the validation set for selecting the optimal hyper-parameters for each algorithm. For each split, the validation set constitutes a random 10% of the training set, thereby excluding any examples from the test scenario or location. Crucially, the division is made on a video basis, ensuring that all clips from a single video are collectively included either in the training or in the validation sets. The performance on the split with the largest training and validation set (**Pl, US-IND**) is used for hyper-parameter optimization.

Baselines. We initially compare our method with the Empirical Risk Minimization (ERM) baseline (Vapnik, 1999), following standard practice in DG research (Carlucci et al., 2019; Gulrajani and Lopez-Paz, 2020). This involves using cross-entropy loss ($\mathcal{L}_c$) without incorporating a generalization objective. Subsequently, we compare our approach against six Domain Generalization (DG) methods, each integrated with $\mathcal{L}_c$ during training.

Most DG methods necessitate domain labels during the training phase. Hence, we supply these labels as needed and denote such methods with an asterisk (*). During testing, all methods rely solely on video clip input, without any domain-specific information. Our baselines, listed chronologically, include:

- CORAL* (Sun and Saenko, 2016): minimizes the distances between means and covariances of video representations from different scenarios, as well as distances between means and covariances from different locations.
- DANN* (Ganin et al., 2016): utilizes two fully connected layers to form an adversarial network predicting the location, alongside a separate adversarial network for scenario prediction.
- MMD* (Li et al., 2018b): similar to CORAL but utilizes MMD distances (Gretton et al., 2012).

- Mixup (Wang et al., 2020c): augments training data through linear interpolations of samples and labels. Unlike CIR, Mixup only considers randomly selected video pairs rather than reconstructing from all batch videos based on visual similarity. Additionally, Mixup alters the output label, whereas CIR retains the original video class label.
- BoDA* (Yang et al., 2022b): aims to minimize distances between domains, similarly to MMD, but with weights assigned based on both domain and class sizes to address imbalance.
- DoPrompt* (Zheng et al., 2022): learns a unique domain prompt for each scenario and location, which is then appended to visual features prior to classification.

Additionally, we provide the average random chance performance across 10 trials.

Implementation details. We utilize SlowFast features (Feichtenhofer et al., 2019), pre-trained on Kinetics (Carreira and Zisserman, 2017), provided with the Ego4D videos (Grauman et al., 2022). The action representation combines three features into a 6912-D vector, following the approach in (Zhou et al., 2018). These features are captured from the action's onset as associated with the narration, midway to the next action, and just before the beginning of the subsequent action. For text features (512-D), we employ the frozen text encoder from the pre-trained CLIP-ViT-B-32 model (Reimers and Gurevych, 2019).

The f encoder consists of two fully connected layers with a hidden dimension of 4096 and an output dimension of 512, featuring a ReLU activation function and a Batch Normalization layer (Ioffe and Szegedy, 2015). The g encoder is also comprised of two fully connected layers, but with a 512 hidden dimension and a ReLU activation function. The dimensions of the query and key embeddings for reconstruction are set at 128.

For all experiments and methods, we use a batch size of 128 and conduct training over 50 epochs with the Adam optimizer (Kingma and Ba, 2014). The learning rate for CIR is set to $2e^{-4}$, with a decay by a factor of 10 at epochs 30 and 40. The coefficients $\lambda_1 = 1$ and $\lambda_2 = 0.5$ are used in Eq. 4.5. Training is completed in 8 hours on a single Nvidia P100 GPU.

4.5.1 Results

Table 4.3 demonstrates that CIR surpasses all prior approaches on every test split, with an improvement of up to 4.9% and an average advantage of 2.1% over the second-best method. When compared to the ERM baseline, CIR achieves an average improvement of 3.4%, and improves by up to 7.7% on the best split. The extent of improvement varies across splits,

	Ga US-PNA	Cl US-MN	Kn IND	Sh IND	Bu US-PNA	Me SAU	Sp COL	Co JPN	Ar ITA	Pl US-IN	Mean
Random	8.00	10.64	9.13	14.36	9.55	13.04	8.35	10.13	9.86	15.68	10.84
ERM	20.75	22.35	18.69	22.14	20.73	23.51	18.97	24.81	22.75	23.29	21.80
CORAL*	22.14	22.55	19.07	24.01	22.18	24.31	19.16	25.36	23.89	25.96	22.86
DANN*	22.42	23.85	19.27	22.89	22.23	23.70	18.64	25.86	23.86	23.28	22.60
MMD*	22.42	23.60	19.66	24.46	22.08	24.64	19.59	25.87	23.84	24.78	23.09
Mixup	21.97	22.21	19.90	23.81	21.45	24.35	19.01	25.90	23.85	24.41	22.69
BoDA*	22.17	22.78	19.62	22.94	21.46	23.97	19.18	25.68	23.92	24.90	22.66
DoPrompt*	21.92	22.77	20.40	23.67	22.75	24.67	18.24	25.04	24.74	25.24	22.94
CIR (w/o text)	23.39	24.52	21.02	26.62	24.64	27.00	19.66	25.42	25.71	30.17	24.81
CIR	**24.10**	**25.51**	**20.46**	**27.78**	**24.93**	**26.83**	**19.75**	**26.34**	**25.67**	**30.94**	**25.23**

Table 4.3 **Top-1 accuracy on ARGO1M.** Best results are in **bold**, and the second-best results are underlined (excluding CIR without video-text association loss, which is greyed out but included for direct comparison to highlight strong performance even without narrations). $*$ indicates that domain labels are required during training.

with the least significant gains observed in the more challenging splits—those with lower ERM results, e.g., (**Kn**, **IND**) and (**Sp**, **COL**).

CIR does not rely on domain labels during training, which is a common strategy for other methods (indicated by $*$ in Table 4.3). Instead, it leverages textual narrations. We also present results for CIR without textual content (i.e., without $\mathcal{L}_{rt}$) or domain labels, showing CIR's robust average performance even with less supervision compared to other strategies.

The second-highest performing method differs across splits, highlighting the problem's complexity and underscoring the necessity of multiple test splits for accurately evaluating domain generalization techniques. Notably, MMD (Li et al., 2018b), a standard DG approach, ranks second best overall, with newer methods finding it challenging to exceed its performance. Techniques that strive to learn domain-invariant visual features, either by matching distributions or through domain prompts, appear to struggle when faced with the scenario shift introduced in ARGO1M. The success of CIR indicates that a reconstruction combined with the use of text narrations offers an effective solution.

4.5.2 Ablations

CIR Ablation. CIR has two reconstruction objectives, and offers three architectural choices for reconstruction, which are ablated in Table 4.4. For the two objectives, the one performing the best differs per split, with the classification reconstructions ($\mathcal{L}_{rc}$) performing better on average (worse results are obtained when it is excluded). Both objectives significantly

	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
CIR (ours)	25.51	**24.93**	26.34	**25.67**	**30.94**	**26.68**
$-\mathcal{L}_{rt}$	24.83	24.80	25.06	25.38	29.50	25.91
$-\mathcal{L}_{rc}$	23.13	23.53	25.87	24.95	26.59	24.81
$-\mathcal{L}_{rt} - \mathcal{L}_{rc}$	22.35	20.73	24.81	22.75	23.29	22.78
$\oplus v$ cross-product	**25.66**	24.84	25.42	25.41	30.67	26.40
$\oplus v'$ learnt att.	22.58	22.55	25.85	24.53	25.35	24.17
$\oplus v = \oplus v'$	23.47	23.33	25.53	24.06	28.74	25.03
$h \neq h'$	24.47	23.12	**26.74**	24.74	27.37	25.29

Table 4.4 **CIR components.** Ablation studies on CIR show the contributions of the two reconstruction strategies and explore alternative design choices, illustrating their influence on the method's effectiveness.

SL	SS	OL	OS	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
✓	✓	✗	✓	25.01	24.86	25.73	**25.99**	30.69	26.46
✓	✓	✓	✗	25.00	25.05	26.07	25.62	30.98	26.55
✓	✓	✗	✗	24.87	24.68	25.77	25.38	30.07	26.15
✗	✓	✓	✓	24.89	**25.13**	26.05	25.80	30.47	26.47
✓	✗	✓	✓	25.22	24.99	26.34	25.84	30.25	26.53
✗	✗	✓	✓	25.17	24.97	**26.36**	25.61	30.31	26.48
✓	✓	✓	✓	**25.51**	24.93	26.34	25.67	**30.94**	**26.68**

Table 4.5 **Effect of masking samples in the support set used for reconstruction.** Columns indicate whether the query can (✓) or cannot (✗) attend to samples from the **S**ame **S**cenario/**L**ocation (**SS**, **SL**) or **O**ther **S**cenario/**L**ocation (**OS**, **OL**) based on the domains they belong to. Note that CIR (bottom) does not use any masking.

outperform the baseline (- $\mathcal{L}_{rc}$ - $\mathcal{L}_{rt}$) without reconstruction. We also ablate other decisions in the reconstruction. Recall that $\oplus v$ is obtained using learnt attention, while $\oplus v'$ utilizes cross-product attention. We show the impact of reversing each of these decisions. Additionally, we found that utilizing the same reconstruction for both ($\oplus v' = \oplus v$) and employing distinct classifiers ($h \neq h'$) yield sub-optimal results.

Attention Masking. CIR reconstructs each clip from others in the batch. On average, a batch contains 11% of videos from the same scenario, 9% from the same location, and 3% from both. We do not limit the samples to attend to, except for avoiding reconstruction from the clip itself. In Table 4.5, we explore possible masks for **S**ame **S**cenario/**L**ocation (**SS**, **SL**) or **O**ther **S**cenario/**L**ocation (**OS**, **OL**). The results indicate that not applying any mask yields the best performance on average, followed by results where the same/other scenario is masked. Masking proves beneficial in certain splits; for instance, excluding samples from different locations enhances performance for (**Ar, ITA**). While we do not

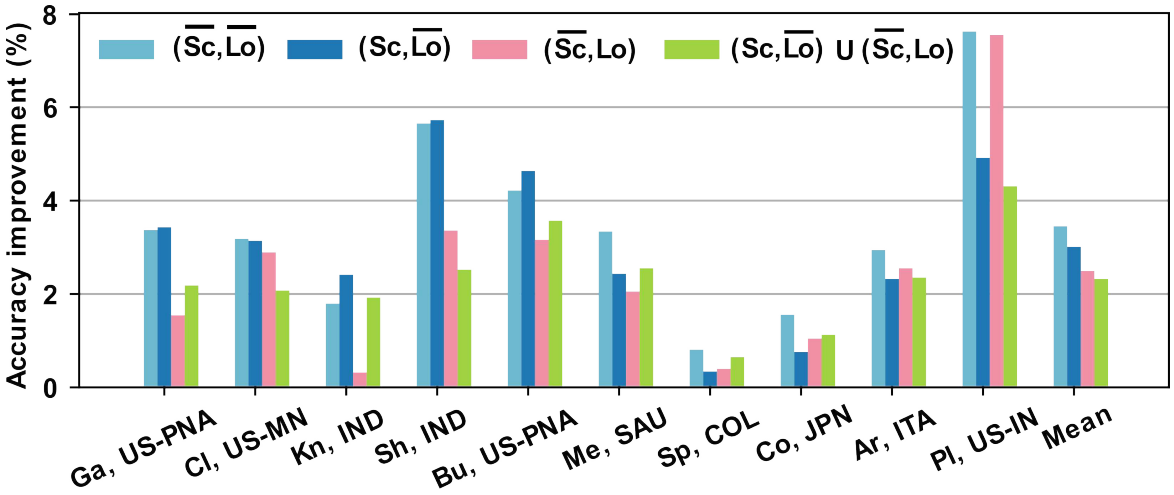


Fig. 4.8 **Effect of scenarios and locations.** Accuracy improvement of CIR over ERM using the same training: (1) neither the test scenario nor location appears in training **(Sc,Lo)**, (2) w/ scenario samples (**(Sc,Lo)**, (3), w/ location samples(**Sc,Lo**)), and (4) w/ both (**(Sc, Lo)**∪(**Sc, Lo**)).

employ masking (thus avoiding the need for domain labels), we highlight its potential when additional information about the domain shift is available.

Effect of scenarios and locations on CIR. Figure 4.8 shows the top-1 accuracy improvements achieved by CIR compared to ERM when both approaches have access to samples from the test scenarios and locations during training. Four scenarios are analyzed: **(Sc,Lo)**, **(Sc,Lo)**, **(Sc,Lo)**, and **(Sc, Lo)**∪**(Sc, Lo)**. CIR exhibits an enhancement over ERM in every scenario and across all splits, with the most significant improvement noted in the most challenging scenario, **(Sc,Lo)**, where both scenarios and locations are not seen during training.

Support-Set Size. In Table 4.6 we show the impact of the batch size on CIR which influences the size of the support set used for reconstruction. CIR is relatively stable over a range of sizes, with slightly worse performance for very small or very large batch sizes.

Text models. We compare the CLIP-ViT-B-32 text encoder with other pre-trained language models in Table 4.7. The results are similar across different language models.

CIR leverages text narrations to mitigate domain shifts. Table 4.8 demonstrates the advantages of this strategy, indicating that simply integrating video-text association into existing methods is not enough. We introduce the text association loss L_{rt}, which acts directly on video representations (i.e., without reconstruction), to current DG methods. We evaluate MMD, which ranks as the second-best performer following CIR and requires domain labels. Additionally, we present results for ERM and Mixup, which do not need domain labels,

	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
16	23.90	22.99	26.04	23.87	28.46	25.05
64	23.89	24.36	**26.54**	24.98	28.97	25.75
128	**25.51**	24.93	26.34	25.67	**30.94**	**26.68**
256	25.00	**24.97**	26.52	**25.96**	30.61	26.61
2048	24.66	24.73	25.48	25.53	30.27	26.14

Table 4.6 **Ablation on batch size.** Effect of varying the batch size on CIR.

LM	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
CLIP-ViT-B-32 (Radford et al., 2021)	**25.51**	24.93	26.34	25.67	**30.94**	**26.68**
all-mpnet-base-v2 (Song et al., 2020)	25.15	25.01	26.30	**25.73**	30.71	26.58
all-miniLM-L6-v2 (Wang et al., 2020b)	25.08	**25.36**	**26.36**	25.45	30.50	26.55

Table 4.7 **Ablation on text models.** Comparison of pre-trained text models.

	T	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
ERM		22.35	20.73	24.81	22.75	23.29	22.78
MMD*		23.60	22.08	25.87	23.84	24.78	24.03
Mixup		22.21	21.45	**25.90**	23.85	24.41	23.56
CIR		**24.52**	**24.64**	25.42	**25.71**	**30.17**	**26.09**
ERM	✓	23.32	23.30	25.84	24.31	27.32	24.82
MMD*	✓	23.69	23.43	25.90	24.27	27.66	24.99
Mixup	✓	23.94	22.94	25.45	24.71	28.52	25.11
CIR	✓	**25.51**	**24.93**	**26.34**	**25.67**	**30.94**	**26.68**

Table 4.8 **Impact of adding text to existing DG methods.** T indicates text supervision. * requires additional domain label supervision.

offering a comparable level of supervision to CIR. Notably, CIR *without text* outperforms other methods *with text*.

Ablation on λ values. We evaluate how CIR results vary as we change λ_1 and λ_2, which weigh $\mathcal{L}_{rt}$ and $\mathcal{L}_{rc}$ respectively. For hyper-parameter selection, we chose the λ_1 and λ_2 values achieving the best results on the validation set (λ_1=1, λ_2=0.5). In Figure 4.9, we plot performance as we vary both λ_1 and λ_2 on the test splits. When λ_1 variations are shown, λ_2 is set to 0.5, and vice-versa. Overall, performance is more sensitive to λ_2 than λ_1. In both cases, we observe a performance drop for lower and higher values.

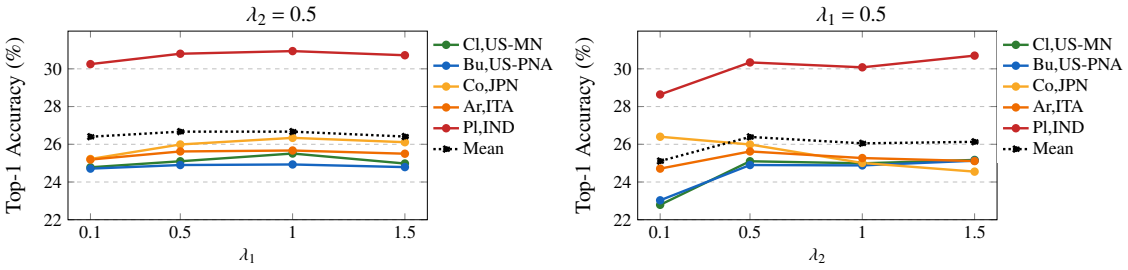


Fig. 4.9 **Ablation on λ values.** Average Top-1 accuracy of CIR, over test splits, as we vary the loss weighting hyper-parameters. Left: Varying λ_1 (left) while keeping $\lambda_2 = 0.5$; as well as varying λ_2 (right) while keeping $\lambda_1 = 0.5$.

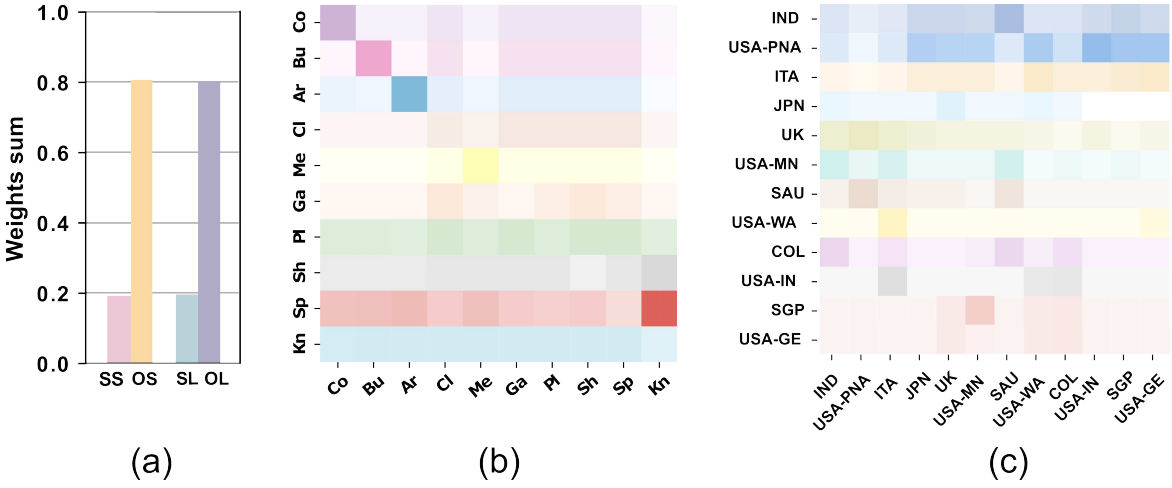


Fig. 4.10 **analysis of attention during reconstruction.** (a) Normalized sum of attention weights over SS, OS, SL, OL. (b) Cross-scenario attention (c) Cross-location attention.

4.5.3 CIR analysis

Figure 4.10 analyzes how videos attend to other videos during the reconstruction-text association process. (a) demonstrates that videos predominantly focus on different scenarios and locations, helping to develop representations that generalize across domain shifts. (b) shows attention between scenarios, with strong self-attention (e.g., cooking) alongside cross-attention (e.g., sport attending to knitting). Some scenarios distribute their attention equally across all scenarios (e.g., playing). (c) depicts attention between locations, where fewer strong entries suggest that knowledge from all locations contribute positively.

Figure 4.11 presents selected examples of our reconstructions during training. It showcases the top-5 support set videos with the highest weights in the reconstruction process (right) in comparison to the query video (left), as identified through CIR (c_{ij}, Section 4.4). CIR is able to attend to samples belonging to other scenarios, other locations, and both. For

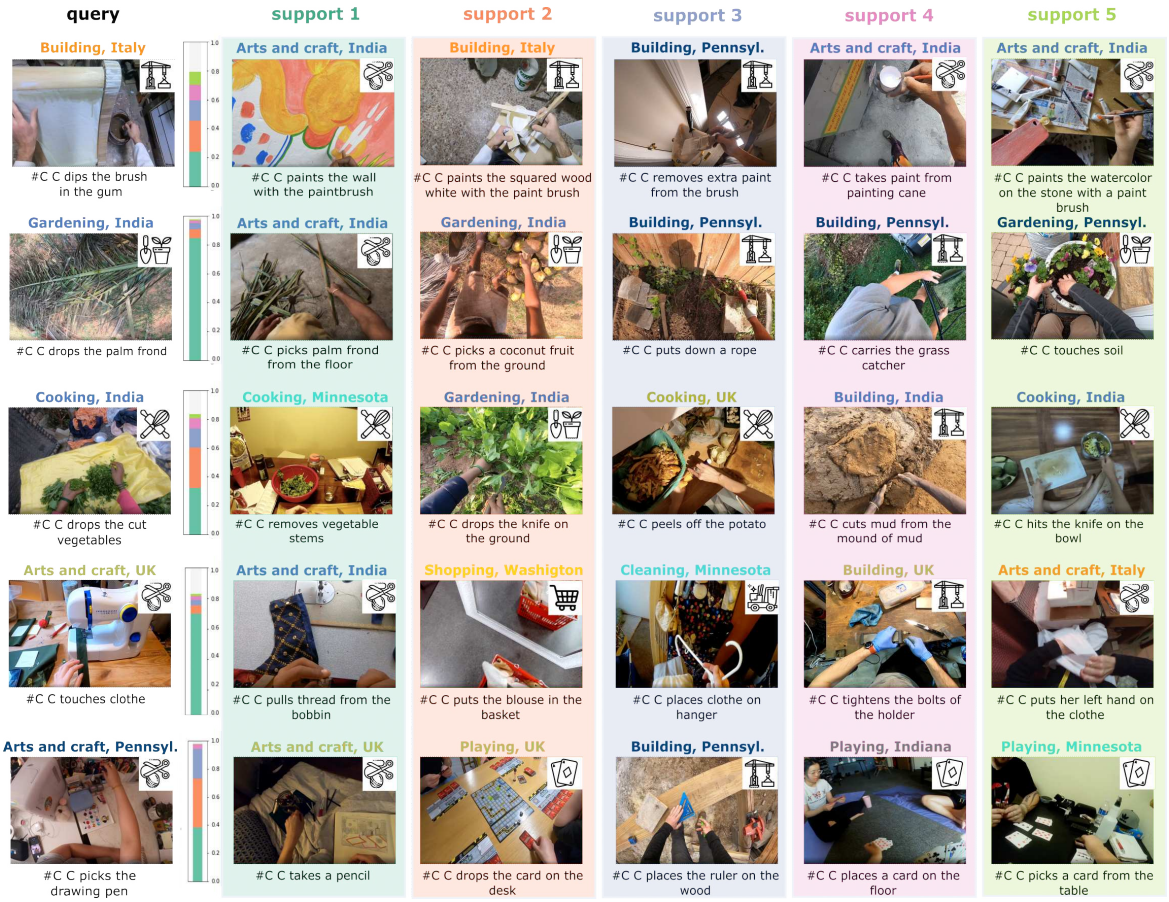


Fig. 4.11 **CIR weights for reconstruction.** Five examples of cross-instance reconstruction from the training set. The query video is shown on the left. For each video, we show its corresponding scenario/location/narration. For each query, the bar shows the score of the j-th support video (colour-matched) with white indicating the sum of the remaining scores from other samples.

instance, in the top row, a painting video from the 'Building' scenario in Italy is reconstructed using examples from 'Arts and Crafts' in India and 'Building' in Italy.

4.6 Conclusion

In this chapter, we introduce the task of Action Recognition Generalization across different scenarios and locations. We hypothesize that it is feasible to learn actions in such a way that they can generalize to new scenarios (e.g., the action 'cut' in cooking could be applied to recognize 'cut' performed by a mechanic) and new locations (e.g., the action 'cut' observed in Italy could be recognized as 'cut' in India). To tackle this new problem, we

introduce ARGO1M a collection of more than 1 million action clips sourced from 73 unique scenario/location combinations.

We propose a generalization method which uses vision and language. It reconstructs a video using samples from various scenarios and locations, aiming for the learned representation to generalize across test splits featuring diverse scenarios and/or locations. Reconstructions are supervised by a classification loss and video-text association loss, enabling the learning of domain-invariant features. CIR consistently outperforms baselines, supported by thorough analysis and ablation studies.

One notable limitation is the dataset's long-tail distribution, which can hinder performance. This imbalance occurs not only at the level of individual actions but also in the diversity of scenarios and locations represented. Future work could focus on addressing these disparities by developing methods that enhance the representation of underrepresented classes. Extending the analysis to zero-shot action recognition tasks could be particularly valuable, especially given the recent advancements in Large Language Models (LLMs). This approach would test the model's capability to recognize actions it has not been explicitly trained to identify, demonstrating its adaptability and potential for broader applicability in real-world scenarios where training data for specific actions might be limited or unavailable. Furthermore, the potential for defining varying hierarchies of actions opens up new avenues for refinement. While our current approach categorizes actions into broad verb macro-categories, there exists the possibility to delve into more fine-grained sub-categories. For example, considering "trimming" as a specific instance within the broader "cut" category could enable a more detailed and precise understanding of actions. This finer granularity could significantly enhance the model's ability to distinguish between closely related actions, contributing to more accurate action recognition.

Chapter 5

Event-Based Data for Egocentric Vision

Chapter 3 and Chapter 4 address cross-domain issues by combining RGB information with traditional modalities such as optical flow, audio, and text. In this chapter, we analyze how a novel modality can be introduced in this context: event data from event-based cameras.

In egocentric vision, RGB sensors are by far the richest source of visual information. However, the performance of RGB-based action recognition models significantly decreases when the training and test data distributions do not match (David et al., 2010). This issue primarily arises from the tendency of appearance-based networks to focus on background cues and object textures, which are often unrelated to the action being performed and can vary greatly across different environments. Consequently, appearance-independent modalities, such as optical flow encoding motion, have emerged as the preferred choice in contemporary egocentric vision systems, as evidenced by the outcomes of recent EPIC-KITCHENS challenges (Damen et al., 2019, 2020, 2021). However, computing optical flow in this context, using algorithms like TV-L1 (Zach et al., 2007), involves solving resource-intensive optimization problems, leading to considerable test-time computation overheads (Crasto et al., 2019).

Event-based cameras, on the other side, have been recognized for their suitability in online settings (Delbruck, 2016; Gallego et al., 2020a). Their high pixel bandwidth minimizes motion blur, and their extremely low latency and power consumption make these innovative sensors especially effective in egocentric scenarios, where fast motion can adversely affect RGB-based systems. Furthermore, since they capture differential information, event sequences can reveal more about the dynamics of a scene than its appearance, presenting a compelling alternative to RGB frames for focusing on motion. Despite these benefits,

previous research has not explored how to leverage their motion sensitivity in egocentric vision, leaving these devices underutilized in such applications.

In this chapter, we introduce N-EPIC-Kitchens, a novel dataset that, for the first time, facilitates the use of event data for egocentric action recognition. Since N-EPIC-Kitchens is derived through event data simulation, we also present an analysis of the sim-to-real domain gap for event-based data, focusing on the simpler task of object recognition. This analysis aims to uncover how effectively models trained on simulated data can generalize to real-world scenarios, providing valuable insights into the transferability and applicability of event-based models in practical applications.

The work presented in this chapter led to the publication of three works:

- Plizzari*, C., Planamente*, M., Goletto, G., Cannici, M., Gusso, E., Matteucci, M., & Caputo, B. (2022). E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 19935-19947).*
  Online Resources: [Paper], [Project page]

- Cannici*, M., Plizzari*, C., Planamente*, M., Ciccone, M., Bottino, A., Caputo, B., and Matteucci, M. (2021). N-ROD: A Neuromorphic Dataset for Synthetic-to-Real Domain Adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 1342-1347).*
  Online Resources: [Paper], [Project page]

- Planamente*, M., Plizzari*, C., Cannici*, M., Ciccone, M., Strada, F., Bottino, A. & Caputo, B. (2021). Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *IEEE Robotics and Automation Letters, 6(4), 6616-6623.*
  Online Resources: [Paper]

5.1 Introduction

Among various sensors, RGB sensors stand out as the most comprehensive source of visual information. Nonetheless, the effectiveness of RGB-based action recognition models significantly diminishes when there is a distribution mismatch between training and testing datasets (David et al., 2010). This problem, often referred to as environmental bias (Kim et al., 2021b; Munro and Damen, 2020a; Planamente et al., 2022b; Sahoo et al., 2021b; Song et al., 2021b), arises from RGB-based networks' reliance on the specific environment where activities are captured. Such dependency hinders their capability to accurately recognize actions in new or unseen settings. The core of this issue lies in the tendency of appearance-based models to overly focus on background elements and object textures, which often do not correlate with the action taking place and can vary widely across different settings. Consequently, modalities that do not depend on appearance, like motion, have gained favor in egocentric vision systems, as demonstrated in recent EPIC-KITCHENS challenge outcomes (Damen et al., 2019, 2020, 2021). Nevertheless, the optical flow utilized in these systems, derived from RGB frames through the solution of intricate optimization problems (e.g., the TV-L1 algorithm (Zach et al., 2007)), entails considerable computational efforts during testing (Crasto et al., 2019).

Event-based cameras, in contrast, have demonstrated particular suitability for online settings (Delbruck, 2016; Gallego et al., 2020a). Their high pixel bandwidth leads to reduced motion blur, while their extremely low latency and low power consumption make these innovative sensors especially advantageous in egocentric scenarios, where rapid movement often affects negatively RGB-based systems. Additionally, because they convey only differential information, event sequences provide more insight into the dynamics of a scene than its appearance, positioning them as a viable alternative to RGB frames for focusing on motion. Yet, despite the benefits, previous research has not explored leveraging their sensitivity to motion in egocentric vision, where these devices remain unused.

As a first effort in this direction, we introduce N-EPIC-Kitchens, a novel dataset that, for the first time, facilitates the use of event data for egocentric action recognition. It is an expansion of the large-scale EPIC-KITCHENS dataset (Damen et al., 2018), which is particularly attractive due to its diversity of environments (kitchens) and the availability of multiple modalities, namely RGB, optical flow, and audio. These characteristics enable an analysis of the previously mentioned environmental bias and a comparison of event data with well-established modalities.

On the proposed N-EPIC-Kitchens, we introduce two approaches to leverage the intrinsic motion characteristics of event data in this context to solve the action recognition task. The first approach, which we call E^2(GO), enriches conventional 2D and 3D action recognition frameworks with modifications at layer level, aiming to exploit the motion-rich attributes of event data. The second approach, E^2(GO)MO, facilitates the transfer of motion signals from optical flow to event data. This transfer is obtained through a teacher-student network, allowing the exhaustive use of the computationally demanding offline TV-L1 flow in the training phase, while bypassing its calculation during test time.

We acquired N-EPIC-Kitchens using the setup proposed in (Munro and Damen, 2020a), which is capable of generating reliable simulated event data. However, this approach gives rise to an open research question: *how well do simulated data generalize to real data?* To address this question, we analyzed the sim-to-real domain gap in event-based data through a simple object classification task. Particularly, we show how standard unsupervised domain adaptation techniques can be used to help models trained on simulated data transfer effectively to real event data obtained from an event-based camera.

We summarize our contributions as follows:

- We introduce N-EPIC-Kitchens, the first event-based egocentric action recognition dataset, which unlocks the possibility to explore event data in this context (Section 5.2.1);
- We propose E^2(GO) and E^2(GO)MO, two event-based approaches tailored at emphasizing motion information captured by event data in egocentric action recognition (Section 5.2.3);
- We benchmark N-EPIC-Kitchens on popular action recognition architectures, showing the performance of event data alone and when combined with traditional RGB and optical flow modalities. We show that event data can surpass RGB in challenging unseen environments and remain competitive in known environments. This suggests that utilizing event data is a feasible alternative and warrants further investigation in this direction (Section 5.2.4).
- We perform an analysis on the sim-to-real domain gap for event based data and show how standard domain adaptation techniques can be used to address it (Section 5.3);

5.2 Event-Based Data for Egocentric Action Recognition

In this section, we introduce event-based data for egocentric action recognition. In Section 5.2.1, we describe the N-EPIC-KITCHENS datasets and our methodology for collecting it. We then present two approaches to leverage the intrinsic motion characteristics of event data in this context (Section 5.2.3). Finally, we benchmark the N-EPIC-KITCHENS dataset using popular action recognition architectures in Section 5.2.4.

5.2.1 N-EPIC-KITCHENS

Thanks to their focus on capturing only changes in the scene, event-based cameras are particularly efficient in egocentric scenarios. They drastically reduce the volume of data that needs to be processed and acquired, prevent motion blur artifacts, and provide fine-grained temporal information. However, so far, only a limited number of datasets have been made freely available (de Tournemire et al., 2020; Gehrig et al., 2021; Hu et al., 2016; Perot et al., 2020). Despite active efforts in the field to increase their availability, as evidenced by the recent release of event-based versions of ImageNet (Kim et al., 2021c; Lin et al., 2021), there are relatively few datasets for human activity recognition currently available. As depicted in Figure 5.1, most of these focus on action or gesture recognition (Amir et al., 2017; Hu et al., 2016; Innocenti et al., 2021; Miao et al., 2019) in controlled settings where both the camera and the background remain static, limiting the use of event-based cameras in this scenario.

To highlight the benefits and potential of event-based cameras in egocentric scenarios, as well as to explore their complementary and equivalent capabilities in comparison to other modalities, we have extended the EPIC-KITCHENS (EK) dataset (Damen et al., 2018) to the event modality. This dataset stands as a comprehensive repository of egocentric videos, showcasing diverse modalities and environments. Drawing on the approach outlined in (Munro and Damen, 2020a), we selected the three kitchens within EPIC-KITCHENS with the highest number of training action instances, designated as D1, D2, and D3 (Figure 5.2). We evaluated performance across the eight most prevalent action categories: ‘put’, ‘take’, ‘open’, ‘close’, ‘wash’, ‘cut’, ‘mix’, and ‘pour’.

In the following sections, we first briefly recall the operating principles of DVS cameras. We direct the reader to Section 4.5 for a more detailed overview. Then, we outline the approach used to generate N-EPIC-KITCHENS and highlight its benefits.

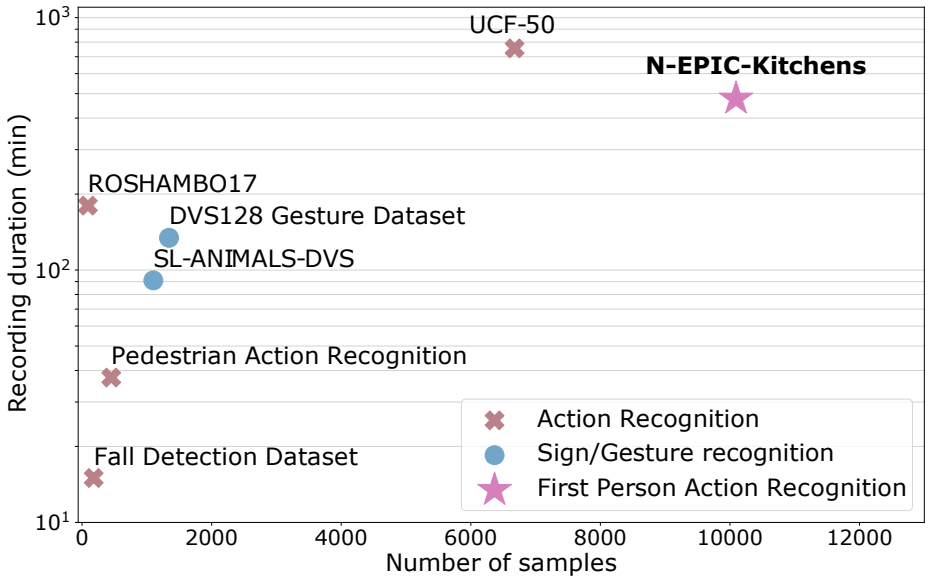


Fig. 5.1 **Dataset comparison.** N-EPIC-KITCHENS *vs* existing event-based action classification datasets in the literature (Amir et al., 2017; Hu et al., 2016; Lungu et al., 2017; Miao et al., 2019; Vasudevan et al., 2020).

Event-Based Vision Data. Pixels in DVS cameras are independent and respond to changes in the continuous log brightness signal, unlike those in a standard RGB camera. An event is a tuple $e_k = (x_k, y_k, t_k, p_k)$, specifying the time t_k, the location (x_k, y_k), and the polarity $p_k \in -1, 1$ of the brightness change (brightness increase or decrease). An event is triggered when the magnitude of the log brightness at pixel $\mathbf{u} = (x_k, y_k)^T$ and time t_k changes by more than a threshold C since the last event at the same pixel, as described by the following equation:

$$\Delta L(\mathbf{u}, t_k) = L(\mathbf{u}, t_k) - L(\mathbf{u}, t_k - \Delta t_k) \geqslant p_k C. \tag{5.1}$$

Therefore, the output from an event camera is a continuous stream of events described as a sequence $\mathcal{E} = \{(x_k, y_k, t_k, p_k) | t_k \in \tau\}$, where τ represents the time interval.

N-EPIC-KITCHENS generation. We utilize ESIM (Rebecq et al., 2018), a recent event camera simulator, to augment the EPIC-KITCHENS dataset with event modality data. Given that videos in EPIC-KITCHENS are limited at 60 frames per second, significantly lower than the microsecond temporal resolution of event cameras, we initially upscale them to a higher frame rate. For this purpose, we employ Super SloMo (Jiang et al., 2018), recognized for its exceptional capability to produce frames at any desired temporal precision. This process is guided by the adaptive sampling strategy outlined in Vid2E (Gehrig et al., 2020), which we adopt for extracting event streams. Subsequently, we apply Voxel Grid (Zhu et al., 2019a), a

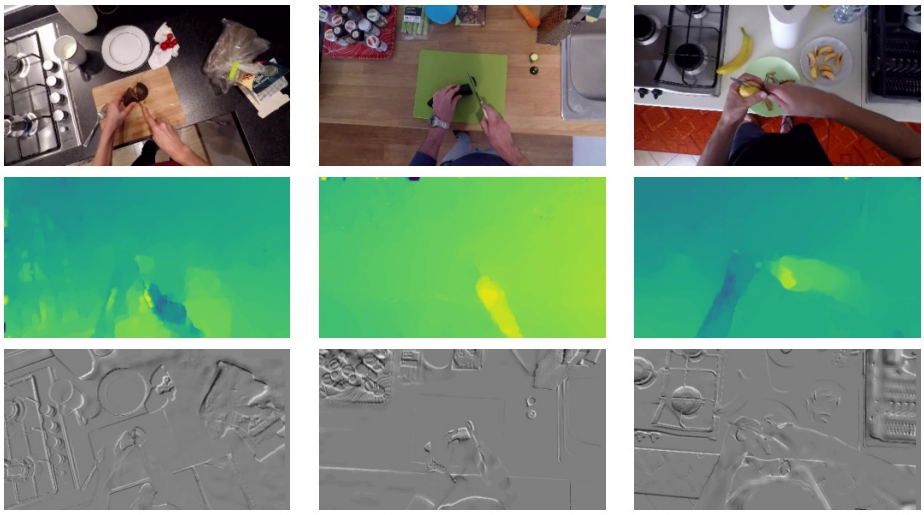


Fig. 5.2 **Multi-modal setting.** RGB (top), optical flow (middle) and Voxel Grid representation (bottom) from the same action ("cut") on the three different kitchens (D1, D2, D3).

technique for encoding event data into frame-like representations, to transform the sparse and asynchronous event data into a tensor format. This enables the application of standard convolutional neural network architectures for learning tasks.

5.2.2 Challenges of evaluating event data

The fundamental challenge in assessing event data for egocentric action recognition lies in its novel application to egocentric vision, unlike other modalities. To establish a benchmark in this area, we evaluate four distinct aspects of event-based modeling. Our evaluation begins by considering performance on both seen and unseen test sets; "seen" refers to performance on the same kitchen used for training, and "unseen" pertains to performance on a different kitchen. We aim to assess both scenarios within our experiments. The performance on seen sets offers insight into the modality's potential maximum effectiveness, while performance on unseen sets tests the model's ability to encode domain-invariant features, thereby indicating its applicability in real-world scenarios. Given that the efficacy of different modalities can significantly vary based on the chosen architecture for processing (Price and Damen, 2019), we benchmark event data using three highly regarded architectures in egocentric action recognition: TSM (Lin et al., 2019), TSN (Wang et al., 2018), and I3D (Carreira and Zisserman, 2017). We utilize a proven method for transforming event streams into a frame-like format, which has been demonstrated to integrate seamlessly with standard

CNNs (Planamente et al., 2021; Stoffregen et al., 2020). Finally, we suggest promoting the modeling of motion features by implementing channel-level attention.

Event Representation. Since event cameras capture scenes through sparse encodings, these encodings must be transformed into intermediate representations for processing. Various representations have been proposed, from bio-inspired (Cannici et al., 2019; Cohen, 2016; Maass, 1997a) to more practical approaches. Frame-like representations are the most widely used methods, as they can be easily integrated with existing network architectures. Among the available options (Cannici et al., 2019, 2020b; Deng et al., 2020b; Gehrig et al., 2019a; Innocenti et al., 2021; Lagorce et al., 2016; Sironi et al., 2018a; Zhu et al., 2019a), we selected Voxel Grid (Zhu et al., 2019a) for its demonstrated superiority in cross-domain applications (Planamente et al., 2021; Stoffregen et al., 2020). This representation generates a B-channel image by dividing time into B distinct intervals:

$$\mathbf{x}^E(x,y,b) = \sum_{k=1}^{N} p_k k_b(b - t_k^*), \tag{5.2}$$

where b represents the channels, t_k^* denotes the timestamps scaled to the range $[0, B-1]$, p_k is the polarity, and $k_b(a) = \max(0, 1-|a|)$.

Backbone Architectures. To evaluate how event data perform across different network designs, we investigate two popular 2D-CNN approaches, TSM (Lin et al., 2019) and TSN (Wang et al., 2018), alongside a 3D-CNN approach, I3D (Carreira and Zisserman, 2017). The first two are based on a 2D-CNN architecture, but while TSN (Wang et al., 2018) primarily utilizes late fusion for temporal modeling, TSM (Lin et al., 2019) employs *shift modules* to facilitate the exchange of channel information across adjacent frames. On the other hand, I3D (Carreira and Zisserman, 2017) is a purely 3D-CNN model that *inflates* its filters and pooling kernels into the temporal dimension. Currently, the literature does not identify a definitive best approach, as different modalities may respond more effectively to one method over the others without a clear pattern.

The importance of motion. Environmental biases are typically addressed in egocentric vision systems by using complementary modalities, often those that do not rely on appearance. Optical flow is usually the best performer in egocentric action recognition tasks (Damen et al., 2018, 2022; Wang et al., 2018), because it (i) focuses on moving content, namely the action being performed, (ii) preserves the edges of moving objects, and (iii) disregards

background information. In this section, we discuss that while event cameras are sensitive to moving edges and capable of ignoring static information, they only partially embody the three key advantages of optical flow mentioned above. Indeed, due to camera movement, these sensors still capture events associated with the background. This observation leads us to consider learning from optical flow to enhance our ability to filter out less discriminative data.

5.2.3 Learning from motion

While a traditional RGB frame captures only static information, frame-based representations utilized for event data also incorporate motion information along the channel dimension (refer to Section 5.2.2). Specifically, each temporal channel encapsulates the motion occurring in the interval between two consecutive frames of the video recording. We introduce two distinct methods to enable standard CNNs to leverage this motion information. The first method, which we name E^2(GO), directly models temporal relationships by integrating channel operations that facilitate motion analysis. The second approach employs a student-teacher strategy, named E^2(GO)MO, aimed at guiding the network to focus on motion features during training through the use of a pre-trained optical flow-based network. We elaborate on these two methodologies in the sections that follow.

E^2(GO): event motion

To allow standard CNNs to capture motion information from event data, we propose two straightforward yet effective architectural modifications. These modifications enhance the ability to extract temporal inter-channel relationships in 2D and 3D CNNs. We refer to these modifications as E^2(GO)-2D and E^2(GO)-3D, respectively.

E^2(GO)-2D. A common approach involves capturing temporal correlations at the video level by modeling dependencies among different frames (Kazakos et al., 2019b; Lin et al., 2019). Event representation uniquely encodes continuous motion, effectively describing micro-movements within the scene. This characteristic motivates us to extend the practice of modeling temporal relations to include learning short-range correlations between event channels.

To achieve this, we leverage *Squeeze and Excitation* modules (Hu et al., 2018) to enhance the attention correlations between channels in 2D CNNs. Given an event volume $\mathbf{x}^E \in$

$\mathbb{R}^{T\times H\times W\times F}$ as input, where T represents the temporal dimension, $H\times W$ denotes the feature map resolution, and F indicates the number of channels, we denote $\mathbf{f}_i^E \in \mathbb{R}^{T\times H_i\times W_i\times C_i}$ as the features extracted from the i-th layer of the network. The first step involves "squeezing" the spatial information content of $\mathbf{f}_i^E$ into a channel descriptor by aggregating features along the spatial dimensions,

$$z_{sq}^E = F_{sq}(f_i^E) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}(f_i{}^E(i,j)) \tag{5.3}$$

where $z_{sq}^E \in \mathbb{R}^{T\times 1\times C}$. Following this is an "excitation" operator, which receives $\mathbf{z}_{sq}^E$ as input to produce a scaling vector $\mathbf{s}$. This vector is used to modulate $\mathbf{x}^E$. The scaling vector $\mathbf{s}$ is derived from $\mathbf{z}_{sq}^E$ through two fully-connected layers, incorporating a bottleneck structure that reduces the channel dimension C to C/r. Subsequently, $\mathbf{s}$ is applied to re-weight $\mathbf{x}^E$, yielding a modified feature vector $\tilde{\mathbf{x}}^E$ that emphasizes discriminative motion features while diminishing less informative ones. Consequently, $\tilde{\mathbf{x}}^E$ captures the dynamic relationships between different temporal channels, effectively modeling their dependencies as a result of a self-attention mechanism on the channel dimension.

E^2(GO)-3D. Similarly, we aim to leverage the capability of 3D CNNs to process temporal information using a 3D kernel. Starting with the same input $\mathbf{x}^E \in \mathbb{R}^{T\times H\times W\times F}$, traditional 3D CNNs apply a 3D convolution across the dimensions of (T,H,W,F), producing an output of shape (T',H',W',C). In our approach, we adapt the 3D convolution operator to work with $\mathbf{x}^E \in \mathbb{R}^{(F\cdot T)\times H\times W\times 1}$ by transposing the channel dimensions onto the temporal axis. This convolution method directly addresses the micro-movements captured across the temporal channels of the event representation, which might be overlooked when processed in the channel dimension alone.

E^2(GO)MO: learning from flow

Our objective is to train a network that utilizes both event data and optical flow data, thereby eliminating the need to estimate optical flow during testing. Given a multi-modal input $X = (X^E, X^F)$, where X^E represents the event modality and X^F represents the flow modality, we denote their respective feature extractors by F^E and F^F, and the extracted features by $\mathbf{f}^E = F^E(\mathbf{x}^E)$ and $\mathbf{f}^F = F^F(\mathbf{x}^F)$. Initially, we train the flow extractor F^F using a cross-entropy loss between the actual action labels $\hat{y}$ and the predicted labels y^F produced by a fully connected layer on top of F^F. Subsequently, we freeze the flow extractor F^F and proceed to train the event stream F^E. This training involves a combination of the standard cross-entropy

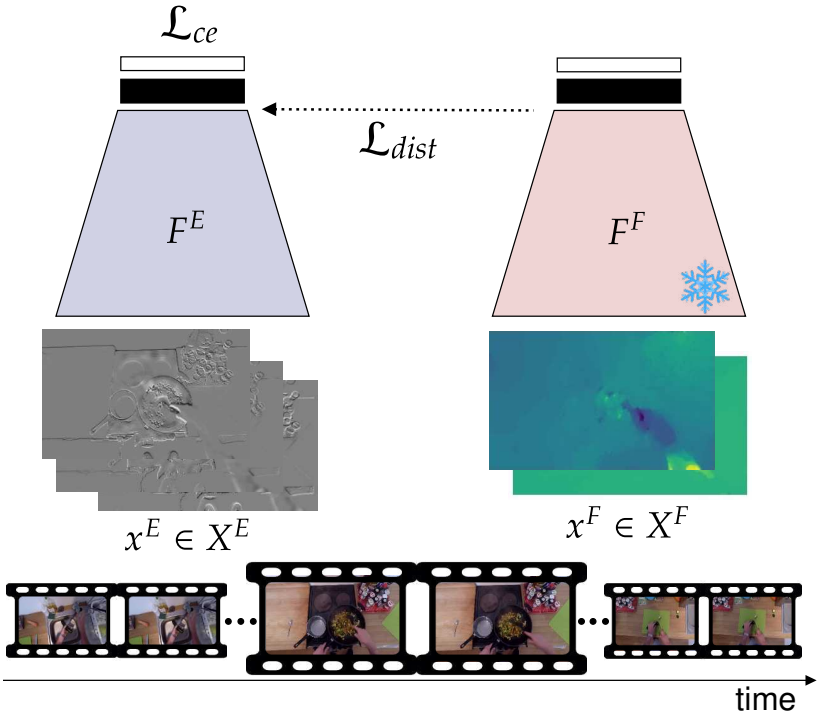


Fig. 5.3 **Illustration of the proposed E^2(GO)MO.** Inputs $\mathbf{x}^E$ from the event modality and $\mathbf{x}^F$ from the flow modality are directed to their respective feature extractors F^E and F^F. Knowledge from the pre-trained (and frozen) teacher stream F^F is transferred to the student stream F^E, which is trained using standard cross-entropy loss.

loss and a *distillation loss*, which is defined as the L_2 norm difference between the features $\mathbf{f}^E$ and $\mathbf{f}^F$:

$$\mathcal{L}_{dist} = \alpha \|\mathbf{f}^E - \mathbf{f}^F\|^2. \tag{5.4}$$

where α is a scaling hyperparameter. This loss encourages the features of the event stream to align with those of the flow stream, compelling F^E to mimic the behaviour of F^F and thereby enabling both to generate similar activations. It is important to note that optical flow data are utilized exclusively during the training phase, and the teacher branch (flow stream) is omitted during inference. This strategy leverages the benefits of the flow modality while effectively circumventing its computational complexity during test. A visual representation of E^2(GO)MO is provided in Figure 5.3.

5.2.4 Experiments

In this section, we first introduce the experimental setup used. We then benchmark event data and evaluate the performance of the proposed E^2(GO) and E^2(GO)MO models. We conclude

the section with a discussion on the findings and a paragraph addressing the limitations of our approach.

Experimental Setup

Input. Experiments involving the I3D model (Carreira and Zisserman, 2017) are carried out by selecting one random clip from the video during training and five equidistant clips during testing, which span the entirety of the video, following the methodology in (Munro and Damen, 2020a). The number of frames in each clip is set to 16 for RGB and optical flow modalities, and 10 for events. For architectures such as TSN (Wang et al., 2018) and TSM (Lin et al., 2019), uniform sampling is employed, involving 5 frames uniformly selected along the video's duration. During testing, the approach involves using 5 clips per video, in line with the procedure described in (Lin et al., 2019). The Voxel Grid representations are bounded between -0.5 and 0.5, and all data modalities are rescaled and normalized to align with the preprocessing requirements of the pretrained networks associated with each adopted architecture. For all modalities, standard data augmentation techniques are applied, as outlined in (Wang et al., 2016).

Implementation and training details. For the I3D model, we opted for the original implementation as detailed in (Carreira and Zisserman, 2017). The TSN and TSM models were constructed using a BN-Inception (Ioffe and Szegedy, 2015) and a ResNet-50 (He et al., 2016) backbone, respectively. In the multi-modal experiments, we employed a classic late fusion strategy, where prediction scores from different modalities are combined through summation, and errors are backpropagated across all modalities. All models were implemented using PyTorch (Paszke et al., 2017). The optimization was carried out using SGD with momentum (Qian, 1999), starting with a learning rate η of 0.01, a weight decay of 10^{-7}, and a momentum μ of 0.9. The networks were trained over 5000 iterations, with a learning rate reduction to $1e{-}3$ at iteration 3000. The experiments were conducted with a batch size of 128 on four NVIDIA Tesla V100 16Gb GPUs. For the distillation loss, the optimal hyperparameter α was determined to be 100. Regarding the evaluation protocol, for *seen* scenarios, we trained on kitchen D_i and tested on the same kitchen ($D_i \rightarrow D_i$), where $i \in 1,2,3$. Performance on *unseen* scenarios was evaluated by training on kitchen D_i and testing on kitchen D_j, with $i \neq j$ and $i, j \in 1,2,3$ ($D_i \rightarrow D_j$).

Model	Voxel ch.	Testing	Seen (%)	Unseen (%)
I3D	3	Clip	53.75	35.90
		Video	**55.54**	**37.52**
TSN	3	Clip	58.81	34.65
		Video	**59.82**	**35.24**
TSM	3	Clip	64.38	37.75
		Video	**65.93**	**38.23**

Table 5.1 **Accuracy on different architectures.** *Mean* accuracy (%) over all $D_i \rightarrow D_j$ combinations on I3D, TSN and TSM on both seen and unseen test sets.

Results

Event Analysis. In Table 5.1 we present the performance of event data across three prominent action recognition architectures (refer to Section 5.2.2). Our findings indicate that utilizing a 3-channel Voxel Grid representation yields the best results, and therefore, we adopted this configuration for all subsequent experiments. When evaluating the performance on both seen and unseen test sets, the TSM model emerges as the most effective, outperforming I3D, which shows slightly inferior results. One possible explanation for I3D's performance is its focus on processing only a limited segment of the video at a time, which limits its ability to capture only local features when trained at the clip level. Conversely, TSM is capable of capturing more global features as it operates on frames spanning the entire video. The sub-optical of TSN is expected, as its method of frame aggregation does not facilitate the modeling of temporal correlations. Thus, unless otherwise stated, we perform video-level analysis and evaluate the proposed approaches on TSM and I3D backbones in all of the following experiments.

Event *vs* RGB. In Table 5.2, we compare the performance of event data against the RGB modality. Results indicate that event data can outperform RGB by up to 3% on unseen test sets. This advantage can be attributed to the literature's findings that appearance-based CNNs are biased towards texture, leading to worse performance across different domains. However, their robustness improves when there is an increased emphasis on shape bias (Geirhos et al., 2018). We hypothesize that the primary reason for the superior performance of event representations is their focus on geometric and temporal information rather than texture variations, making them more invariant to domain changes. This principle also holds for seen tests, where RGB-based networks tend to overfit by relying on domain-specific features. Note that, until this point, event modalities were considered to lag behind RGB images in purely visual tasks, as evidenced by the recent release of the N-ImageNet benchmark (Kim

Modality	Model	D1	D2	D3	D1→ D2	D1→D3	D2→D1	D2→D3	D3→D1	D3→D2	Seen (%)	Unseen (%)
RGB	I3D	53.67	61.12	60.70	34.50	35.70	34.94	36.46	33.93	38.37	**58.49**	35.65
Event	I3D	50.32	58.33	57.99	37.27	39.12	32.98	36.52	35.68	43.56	55.54	37.52
Event	E^2(GO)-3D	50.52	62.99	60.11	38.07	38.71	35.02	38.49	36.73	45.53	57.87	**38.76**
RGB	TSM	61.61	77.08	75.75	37.39	32.49	34.28	38.99	34.43	38.25	**71.48**	35.97
Event	TSM	56.86	72.43	68.49	28.73	34.00	37.09	42.30	42.27	45.02	65.93	38.23
Event	E^2(GO)-2D	56.58	70.03	69.60	34.98	35.16	38.21	47.80	41.71	44.13	65.40	**40.33**

Table 5.2 **E^2(GO) results.** Accuracy (%) with respect to RGB using both I3D and TSM frameworks is presented across all shifts, denoted by $D_i \rightarrow D_j$, indicating training on D_i and testing on D_j, with D_i signifying training and testing on the same dataset. The top performances for both seen and unseen, for each backbone, are in **bold**.

et al., 2021c), where the highest-scoring event architecture achieved only 48.94% accuracy, significantly lower than the greater than 90% accuracy achieved by RGB models (Dai et al., 2021; He et al., 2022; Pham et al., 2021; Zhai et al., 2022). In our study, however, we demonstrate that event data can not only outperform RGB in challenging unseen scenarios but also compete effectively in seen ones. This highlights the significant potential of event data in enhancing egocentric vision applications.

E^2(GO). In Table 5.2, we detail the performance of E^2(GO)-2D and E^2(GO)-3D. These modifications prove particularly advantageous on unseen test sets, as they are designed to improve temporal correlations. This enhancement enables the network to highlight motion features that are informative for the action being performed while de-emphasising those that do not correlate with the action. E^2(GO)-3D improves by up to 2% on the seen test set, whereas E^2(GO)-2D achieves results that are on par with the baseline TSM model. This discrepancy can be attributed to the inherent characteristics of 2D CNNs, which, as frame-based techniques, depend significantly on visual signals. Although these signals can be detrimental on different environments, they may prove beneficial in seen scenarios. Conversely, the I3D model is innately more responsive to temporal correlations. By extending its capacity for temporal reasoning to include micro-movements, it becomes more effective in identifying discriminative features relevant to the action. This capability results in higher accuracy, even when testing in the same environment, demonstrating the potential of integrating enhanced temporal dynamics into action recognition models.

Multi-modal analysis. Table 5.3 demonstrates the synergistic effects of combining the event modality with RGB and optical flow data in action recognition tasks. When integrated with RGB data, the combination results in an improvement of up to 7% on seen test sets

Model	**Streams**	**Pretrain**	**Seen (%)**	**Unseen (%)**
I3D	Event	Kinetics-400 (R)	55.54	37.52
E^2(GO)-3D	Event	Kinetics-400 (R)	57.87	38.76
TSM	Event	ImageNet	**65.93**	38.23
E^2(GO)-2D	Event	ImageNet	65.40	**40.33**
I3D	Event+RGB	Kinetics-400 (R)	59.12	38.13
E^2(GO)-3D	Event+RGB	Kinetics-400 (R)	61.23	**41.85**
TSM	Event+RGB	ImageNet	71.88	39.92
E^2(GO)-2D	Event+RGB	ImageNet	**72.42**	40.61
I3D	Event+Flow	Kinetics-400 (R)	60.48	44.47
E^2(GO)-3D	Event+Flow	Kinetics-400 (R)	62.66	45.86
TSM	Event+Flow	ImageNet	72.26	46.89
E^2(GO)-2D	Event+Flow	ImageNet	**72.87**	**49.23**
I3D	RGB+Flow	Kinetics-400 (R)	62.07	44.56
TSM	RGB+Flow	ImageNet	**75.08**	**45.66**

Table 5.3 **Multi-modal results.** Accuracy results (%) of the event modality when used in combination to stardard RGB and optical flow. In **bold** the best result for each modality combination.

and 3% on unseen ones. However, the most significant performances are observed when event data is paired with optical flow, with improvements reaching up to 7% on seen domains and 9% on unseen ones. This result suggests that although both event data and optical flow capture motion information, optical flow is more focused on the motion-relevant aspects, often overlooking scene or object affordances. In contrast, event data retains valuable information about object shapes, as illustrated in Figure 5.2. Therefore, combining event data with optical flow appears to be more effective than pairing it with RGB, especially in unseen domains where RGB's reliance on appearance can be a disadvantage. Moreover, this combination outperforms the conventional RGB+Flow approach, indicating that standard event representations, which do not emphasize appearance features as strongly as RGB does, can offer distinct advantages in action recognition tasks, particularly in enhancing generalizability and robustness across different environments.

E^2(GO)MO. In Table 5.4 we present the performance of E^2(GO)MO in comparison to an RGB-based TSM, which, according to our previous analyses, emerges as the most robust architecture. To support our hypothesis that the proposed distillation technique is particularly beneficial for leveraging motion features, we apply the same distillation mechanism to an RGB-based stream, denoted in Table 5.4 as RGB+$\mathcal{L}_{dist}$. Both the event and RGB streams show performance improvements on unseen test sets (by +5.3% and +3%, respectively),

Method	Model	D1	D2	D3	D1→D2	D1→D3	D2→D1	D2→D3	D3→D1	D3→D2	Seen (%)	Unseen (%)	Mean (%)
RGB	TSM	61.61	77.08	75.75	37.39	32.49	34.28	38.99	34.43	38.25	71.48	35.97	53.73
RGB + $\mathcal{L}_{dist}$	TSM	63.36	79.47	77.97	38.61	35.73	39.36	41.09	34.76	49.68	**73.60**	39.87	56.73 ▲+3
RGB + Flow	TSM	66.97	79.69	78.58	43.76	43.76	45.80	47.13	45.44	48.09	75.08	45.66	60.37
Event	TSM	56.86	72.43	68.49	28.73	34.00	37.09	42.30	42.27	45.02	65.93	38.23	52.08
Event	E^2(GO)-2D	56.58	70.03	69.60	34.98	35.16	38.21	47.80	41.71	44.13	65.40	40.33	52.87
Event	E^2(GO)MO-2D	61.38	75.83	75.08	39.77	37.19	44.71	51.03	47.01	53.73	70.76	**45.57**	**58.17** ▲+5.3
Event + Flow	E^2(GO)-2D	65.11	77.58	75.91	42.12	41.80	48.20	53.50	51.85	57.91	72.87	49.23	61.05

Table 5.4 **E^2(GO)MO results.** Accuracy (%) of E^2(GO)MO w.r.t. the baseline on events (TSM) and E^2(GO)-2D. We compare E^2(GO)MO with the same approach on RGB to validate the choice of combining event and flow. In **bold** the best uni-modal, underlined the best multi-modal.

reinforcing the significance of motion information in real-world scenarios. However, the E^2(GO)MO model benefits significantly more from the distillation loss $\mathcal{L}_{dist}$ than the RGB stream does, suggesting that event data encapsulates more motion-rich features compared to RGB streams. This observation validates our premise regarding the motion-centric nature of event data. Additionally, we compare these models against their respective multi-modal benchmarks, which incorporate offline-computed optical flow during prediction—specifically, RGB+Flow and E^2(GO)+Flow. While neither model achieves its theoretical upper bound performance, E^2(GO)MO comes closer to matching the E^2(GO)+Flow benchmark and even surpasses the multi-modal performance of RGB+Flow. This outcome further advocates for the utility of event data, highlighting its advantages in egocentric vision applications, especially when the goal is to enhance motion understanding without relying solely on standard RGB information.

Event *vs*. Optical Flow. Figure 5.4 illustrates the trade-off between accuracy and average time per frame at test time on both seen and unseen data. We evaluate performance using two methods of optical flow computation: the TV-L1 algorithm, computed offline as described in (Zach et al., 2007), and the flow extracted from PWC-Net (Sun et al., 2018a), the latter being among the most efficient end-to-end CNN models for flow estimation, striking a favorable balance between time efficiency and accuracy. For these calculations, we utilize a NVIDIA Titan RTX GPU and report on both the computation time of the inputs and the forward pass time, excluding the time for data access. Additionally, we delineate the range within which real-time action recognition is feasible, adopting the frame (sampling) rate threshold from (Song and Godøy, 2016), which is deemed adequate for a motion tracking system. The analysis reveals that while TV-L1 achieves higher accuracy, it does so at the expense of a substantial extraction time of 488 ms, making it impractical for online scenarios. On the other hand, when optical flow is estimated in real-time using PWC-Net, there is a significant

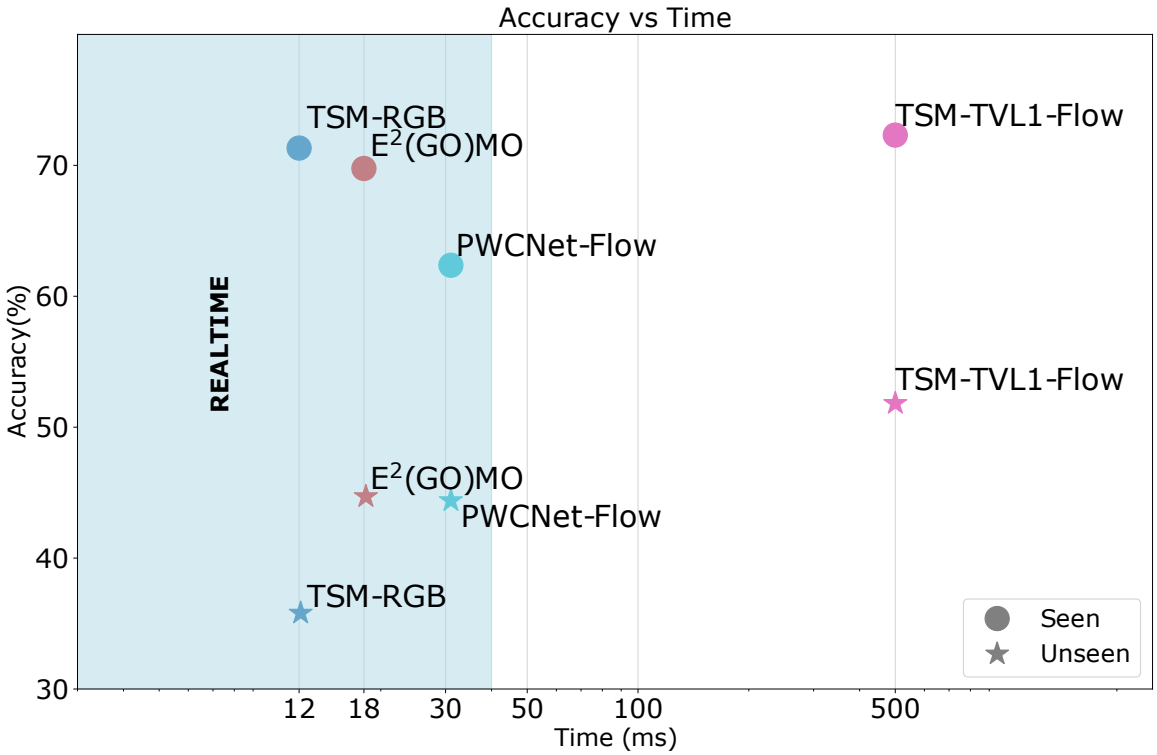


Fig. 5.4 **Accuracy *vs* time** of RGB modality, E^2(GO)MO, estimated PWCNet optical flow and TV-L1 optical flow on seen and unseen scenarios for one clip evalutation.

drop in performance (by up to 10% on seen tests and 8% on unseen tests). Furthermore, the use of PWC-Net introduces the need for an additional network, thereby increasing the total number of parameters (approximately 40M) and necessitating an extra stage of fine-tuning. In contrast, our approach, which does not require the computation of flow at test time, allows us to fully leverage the precision of more accurate optical flow methods during the distillation process. As a result, E^2(GO)MO, despite not explicitly utilizing flow during inference, still manages to outperform PWC-Net in seen tests (by up to 6%) and matches its performance on unseen tests. This outcome underscores the effectiveness of our method in balancing accuracy and computational efficiency, making it a compelling option for real-time action recognition in egocentric vision applications.

Discussion and limitations. The simulation of event camera data introduces an inevitable sim-to-real domain shift, as current methods cannot perfectly replicate the behaviors of actual event cameras (Planamente et al., 2021; Stoffregen et al., 2020). Despite this limitation, research has demonstrated that simulated events possess sufficient robustness to generalize effectively to real-world scenarios (Gehrig et al., 2020; Planamente et al., 2021; Stoffregen et al., 2020). As we introduce event data into the domain of egocentric action recognition for

the first time, our goal is to provide a direct comparison with established benchmarks in the literature (Damen et al., 2018, 2022; Fathi et al., 2012b) and to position the event modality as a competitive alternative to traditional modalities. This strategic choice motivates our decision to simulate event data instead of creating a new first-person dataset from the ground up.

In the next section, we delve into how simulated event data can be effectively adapted to real-world scenarios using domain adaptation techniques. We aim to showcase the potential of these techniques to mitigate the sim-to-real domain shift, ensuring that models trained on simulated data can perform reliably on real event data. This exploration is set to highlight the adaptability and applicability of simulated event data in practical egocentric vision tasks.

5.3 Sim-to-Real Gap in Event-Based Data

Recently, new learning approaches that utilize standard computer vision algorithms on event data have achieved competitive results compared to traditional methods (Gehrig et al., 2019b; Maqueda et al., 2018). However, training these state-of-the-art deep learning algorithms demands a substantial amount of data, a requirement that is currently limited by the novelty and high cost of neuromorphic cameras. A practical solution to this data scarcity issue is the use of event camera simulators (Rebecq et al., 2018), which can produce reliable simulated event data. Yet, this solution prompts an important research question: how well do simulated data generalize to real data? This challenge has been recently addressed in part by (Gehrig et al., 2020) and (Stoffregen et al., 2020), who proposed methods to narrow the sim-to-real gap by tweaking simulator parameters. These adjustments occur at the input level during the data simulation phase, indicating a strategic approach to making simulated data more reflective of real-world conditions.

While (Gehrig et al., 2020) and (Stoffregen et al., 2020) address the sim-to-real gap by focusing on the generation of event data, we approach the issue from a domain adaptation perspective, viewing it as a domain-shift problem. Unlike the well-known Synth-to-Real shift, which concerns the visual appearance differences between rendered RGB images (ble) and real RGB images, the challenge here involves a different kind of shift. Specifically, the gap arises from the differing distributions of events in response to local brightness changes. Simulators often overlook certain non-idealities that are inherent to real event cameras, such as the minimum intensity change threshold required to trigger an event or the refractory period of event pixels, which can vary across different event cameras.

In this section, we demonstrate how Unsupervised Domain Adaptation (UDA) techniques can effectively bridge the Sim-to-Real gap for event cameras by aligning the feature distributions between the simulated source domain and the real target one. This alignment allows neural networks to leverage both simulated data and real, unlabeled events during training. We extend our analysis to the Synth-to-Real gap by comparing synthetic rendered images and real images, each paired with corresponding simulated events, to investigate how the simulated event modality is affected this shift and benefits from UDA techniques. To facilitate this analysis, we introduce a specialized multi-modal dataset, N-ROD, which includes real event data captured with an event camera, and data generated through simulation, paired with real and synthetic images from ROD (Loghmani et al., 2020). We name our approach DA4Events (DA4E), and illustrate the different domain shift we analyze in Figure 5.5.

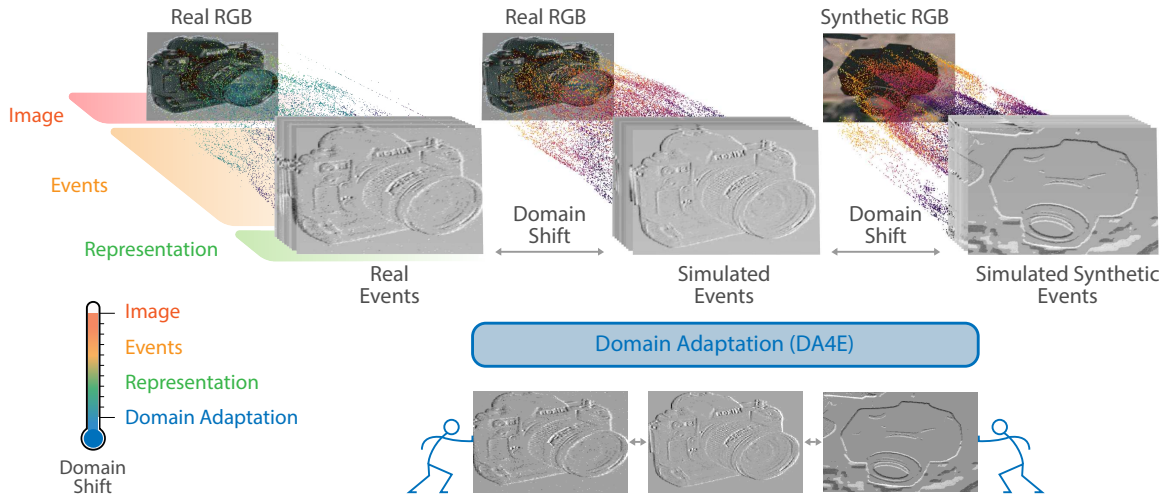
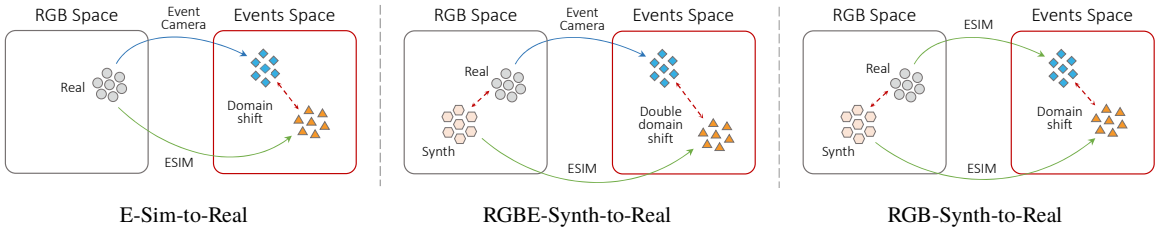


Fig. 5.5 **Sim-to-Real gap in event-based cameras.** DA4Events exploits unsupervised domain adaptation techniques to solve this problem by acting at feature level. *How else simulated events can be used?* We propose to use events in a real context, exploiting the complementarity with RGB data to improve networks robustness.

5.3.1 Formulation

In the context of Unsupervised Domain Adaptation (UDA), our objective is to train a model on a source domain $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, which contains N_s labeled samples with labels from a known label space $\mathcal{Y}^s$, such that it also performs well on a target domain $\mathcal{T} = \{x_i^t\}_{i=1}^{N_t}$, comprising N_t unlabeled samples from an unknown label space $\mathcal{Y}^t$. We operate under two primary assumptions: (i) the source and target domains exhibit different distributions, denoted as $\mathcal{D}_s \neq \mathcal{D}_t$, and (ii) both domains share the same label space, meaning $\mathcal{Y}s = \mathcal{Y}_t$. The final goal is to align the source and target domain distributions by leveraging UDA techniques outlined in Section 5.3.2, thereby enhancing the model's ability to transfer from the source to the target domain.

To demonstrate that the proposed approach is general, we focus on examining different domain gaps affecting event generation. These settings impact in different ways the two domain distributions $\mathcal{D}_s$ and $\mathcal{D}_t$. We start by analysing the *E-Sim-to-Real* and *RGBE-Synth-to-Real shifts*. While the first solely accounts for discrepancies in event generation, the second entails a dual shift that also includes variations in the RGB space. Additionally, we consider a simple variant termed *RGB-Synth-to-Real*, where the event generation process is identical across both domains, with the sole discrepancy arising from a shift in the RGB space. An overview is presented in Table 5.5, and a visual representation can be found in Figure 5.6. In the next section, we provide a more detailed description of these shifts.



E-Sim-to-Real RGBE-Synth-to-Real RGB-Synth-to-Real

Fig. 5.6 **Domain shifts.** Visualization of the three domain shifts studied in this chapter. Clusters of symbols represent data in the RGB / events space, while arrows indicate event generations through simulation (ESIM) or through an event camera (Event Camera).

		Source		Target	
Setting	Dataset	RGB	Event	RGB	Event
E-Sim-to-Real	N-Caltech101		ESIM(RGB_{real})		EvCamera(RGB_{real})
RGB-Synth-to-Real	ROD	Synth	ESIM(RGB_{synth})	Real	ESIM(RGB_{real})
RGBE-Synth-to-Real	N-ROD	Synth	ESIM(RGB_{synth})	Real	EvCamera(RGB_{real})

Table 5.5 **Comparison between the different settings.** We indicate as ESIM(·) the events obtained through simulation (Rebecq et al., 2018) from either synthetic or real RGB images, and with EvCamera(·) those obtained using a real event camera. We indicate Sim-to-Real and Synth-to-Real in different colors, and highlight the corresponding shift in the right side of the table using the same color.

E-Sim-to-Real shift. In this shift, our focus is solely on the differences within the event generation process. Simulated events in the source domain, $\mathcal{E}^s_{sim}$, are generated from an RGB dataset using an event simulator, i.e., $\mathcal{E}^s_{sim}$ = ESIM(RGB^s_{real}), and paired with real events in the target domain captured from the same RGB images using an actual event camera device, i.e., $\mathcal{E}^t_{real}$ = EvCamera(RGB^t_{real}). An example of this shift is shown in Figure 5.7. Some aspects of this shift have been partially addressed through the refinement of hyperparameter selection in simulation methodologies (Gehrig et al., 2020; Stoffregen et al., 2020). We utilize this scenario to examine the effectiveness of unsupervised domain adaptation (UDA) methods in managing variations in the event generation process.

RGBE-Synth-to-Real Shift. This shift examines the combined effect of RGB rendering and event simulation. Similar to the previous scenario, the target domain consists of event streams captured with a real camera, namely $\mathcal{E}^t_{real}$ = EvCamera(RGB^t_{real}). However, in this case, the source domain comprises simulated events derived from synthetic renderings, $\mathcal{E}^s_{sim}$ = ESIM(RGB^s_{synth}). This particular configuration represents a double shift as the change in event generation (ESIM → EvCamera) is coupled with that in RGB images (RGB^s_{synth} →

(a) RGB image (b) Real events (c) Simulated events

Fig. 5.7 **Real *vs* simulated events.** Real and simulated events (voxel grid (Zhu et al., 2019a)) on a Caltech101 sample.

RGB^t_{real}). We demonstrate that our methodology is adaptable enough to manage this shift without any modifications to the general framework.

RGB-Synth-to-Real Shift. Finally, we explore a simplified scenario of the previous setting where the shift in event generation is neutralized by simulating events for both the source and target domains. The source events are thus simulated as $\mathcal{E}^s_{sim} = \text{ESIM}(RGB^s_{synth})$, and the target events as $\mathcal{E}^t_{sim} = \text{ESIM}(RGB^t_{real})$. We employ this approach to investigate how variations in the RGB domain impact on performance.

5.3.2 DA4Event: Domain Adaptation for Event Data

We consider a general multi-modal framework where both domains provide paired images and events, i.e., $(RGB^s, \mathcal{E}^s)$ in the source domain and $(RGB^t, \mathcal{E}^t)$ in the target domain. Images can either be captured by a real camera or obtained through rendering (Blender, 2018), while events can be generated from an actual event-based camera or through event simulation (Rebecq et al., 2018) starting from one of the two image domains. We transform the event streams $\mathcal{E}^s$ and $\mathcal{E}^t$ into multi-channel event representations $\mathcal{R}^s_\mathcal{E}$ and $\mathcal{R}^t_\mathcal{E}$, employing techniques from the literature. When feasible, we use a window-based computation method to divide the event stream into consecutive bins, extracting a representation from each to improve performance, as evidenced by prior studies. This yields representations $\mathcal{R}^{s,t}_\mathcal{E} \in \mathbb{R}^{H\times W\times F}$, where the number of features F depends on the chosen representation and the number of bins used.

These multi-channel event representations are fed into a feature extractor $\mathcal{F}_\mathcal{E}$, compatible with the UDA methods discussed earlier and shared between the source and target domains. The source domain samples' extracted features are then processed by a classifier $\mathcal{G}$ and a

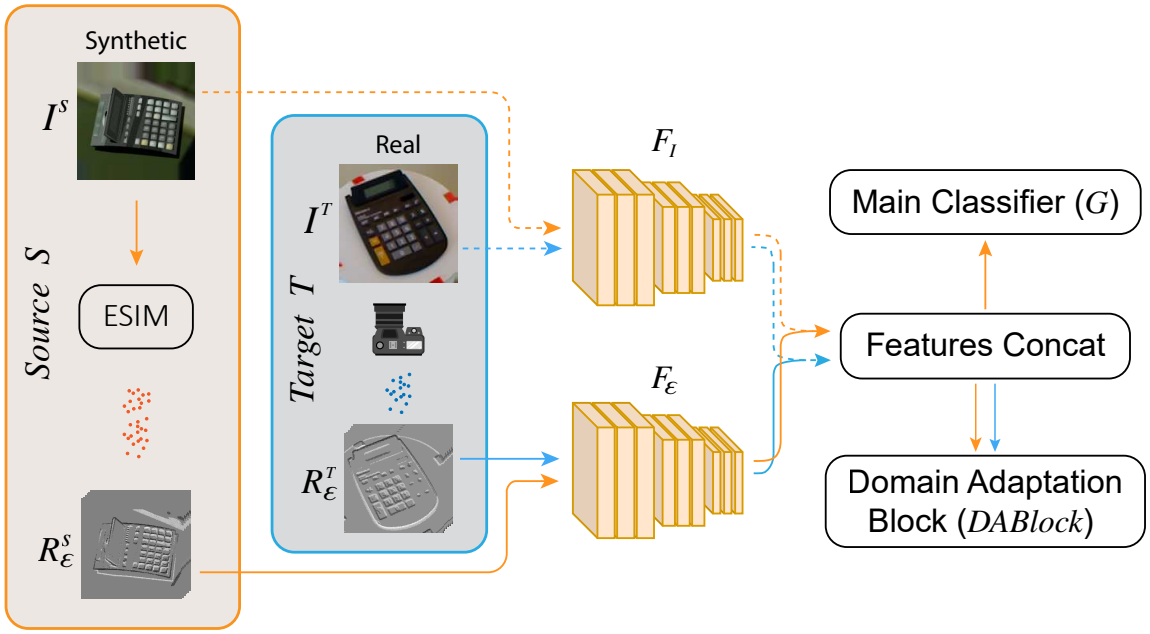


Fig. 5.8 **Multi-modal DA architecture.** Data coming from the source and target domains are processed separately during training. Source, labelled, data is used for supervised classification in $\mathcal{G}$, while both target and source data are fed to the *DABlock*. Features are extracted from each modality using different extractors $\mathcal{F}_I$ and $\mathcal{F}_\epsilon$, shared across domains, and then concatenated before prediction. The dashed data path is finally removed, along with features concatenation, when just the event modality is used.

domain adaptation block (DABlock), which also integrates target domain features for adaptation. This DABlock encompasses the domain adaptation techniques detailed in a previous section. Throughout training, the DABlock aims to minimize the primary classification loss $\mathcal{L}_y$, based on $\mathcal{G}$'s predictions, alongside an auxiliary domain adaptation loss $\mathcal{L}_{DA}$, thereby facilitating the regularization of the model towards effective domain adaptation.

MV-DA4Event: a Multi-View Approach. A common method for handling event data involves aggregating the event stream $\mathcal{E} = \{e_i = (x_i, y_i, t_i, p_i)\}_{i=1}^N$, which captures the spatio-temporal dynamics of a scene over a temporal interval T, into a frame-based representation $\mathcal{R}_\mathcal{E} \in \mathbb{R}^{H\times W\times F}$. This transformation facilitates the processing of event data using conventional convolutional neural networks (CNNs). Unlike standard RGB images that encode solely spatial (static) information through the R, G, B channels, these frame-based event representations encompass temporal information. This is achieved by dividing the event sequence into multiple intervals (or bins), similar to frames in a video sequence, to preserve temporal detail. For example, in saccadic motion, a technique often employed to capture event data from stationary planar images, the channels represent the camera's response to different movement

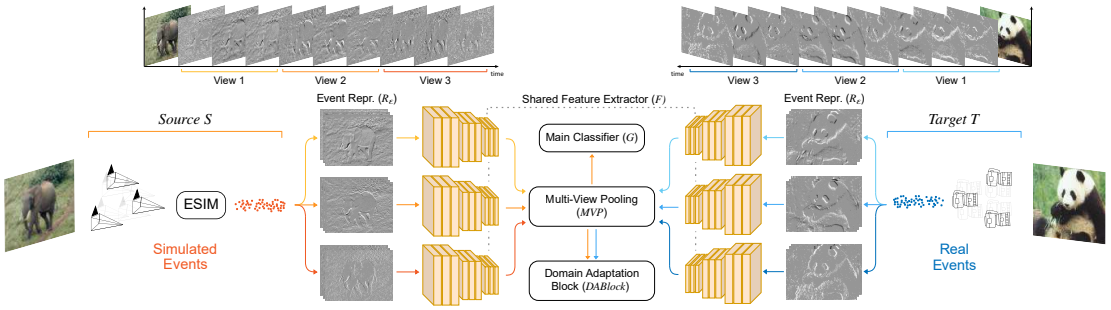


Fig. 5.9 **MV-DA4E architecture.** Top shows the process of extracting an event representation, taking voxel grids (Zhu et al., 2019a) and three views as an example, while bottom details the proposed multi-view architecture (MV-DA4E). Two unpaired random batches from source and target domains are sampled and processed separately during training. When the multi-view approach is not used (DA4E), event representations are fed as a single multi-channel tensor to the feature extractor $\mathcal{F}$, and multi-view pooling is removed. Notice that only source (labelled) data are fed to the classifier $\mathcal{G}$, while both target and source data are fed to the DABlock.

directions. Consequently, each temporal channel offers a distinct perspective of the observed object, emphasising various features.

In computer vision and event-based processing, it is common to initialize CNNs with weights pre-trained on ImageNet. However, for a k-channel representation, where $k \neq 3$, the usual method involves replacing the first convolutional block with a new one and training it from scratch. This approach might not only limit the utilization of the pre-trained model but could also be detrimental in cross-domain scenarios. Literature suggests that the initial layers of a network are often most impacted by domain shift, leading a network trained from scratch on these layers to specialize too much on the source domain, hindering generalization to the target domain. Conversely, transferring pre-trained layers allows the network to leverage robust low-level features, enhancing adaptability.

Motivated by these considerations, we propose a *multi-view* strategy to preserve the first pre-trained convolutional layer. This involves transforming the multi-channel event representation into three-channel images, or *views*, resulting in a representation $\widetilde{\mathcal{R}}_{\mathcal{E}} \in \mathbb{R}^{H \times W \times \lceil F/3 \rceil \times 3}$. A *multi-view* network has been specifically crafted, where each *view* is independently processed by a feature extractor $\mathcal{F}$. The collected set of features is then merged using a late-fusion approach within a Multi-View Pooling (MVP) module, which applies average pooling to produce a feature vector in $\mathbb{R}^{F_{out}}$. This vector is subsequently utilized throughout the remaining segments of the network. Given that the initial layers of the network are more prone to domain-specific influences while the later layers encapsulate more task-specific knowledge, we hypothesize that merging the different views in the network's final stages, rather than at the beginning, fosters enhanced generalization capabilities.

Network architecture. In Figure 5.9 we present the architecture of our proposed network. Events are initially generated using the ESIM simulator in the source domain and directly captured from an event-based camera in the target domain. These events are divided into B temporal bins, from which a sequence of event representations is derived, resulting in a multi-channel volume $\mathcal{R}_{\mathcal{E}}$ with channels that are a multiple of 3. These representations are then organized into group views, specifically, 3-channel frames that are interpreted as images. These frames are processed in parallel through a shared ResNet feature extractor $\mathcal{F}$. The output features from this process are subsequently merged in the Multi-View Pooling (MVP) module, which conducts average pooling both spatially and across the views for features within the same domain, producing two distinct feature vectors for each domain. The features from the source domain are utilized in $\mathcal{G}$ for making final predictions and in the Domain Adaptation Block (DABlock), along with features from the target domain, to facilitate domain adaptation. It is important to note that during training, two completely random batches of source and target samples are selected without any matching constraints between them.

UDA Algorithms

In this section we give a brief overview of the UDA methods applied within the *DABlock* of our architecture.

Gradient Reversal Layer (GRL). The concept of GRL (Gradient Reversal Layer) involves integrating Domain Adaptation (DA) into the feature learning process. This goal is accomplished by simultaneously optimising the label predictor and a *domain classifier*, which is tasked with determining whether a sample originates from the source or the target domain (Ganin and Lempitsky, 2015b). The training process is designed to deceive the domain classifier by maximising its loss through the use of a gradient reversal layer, thereby promoting the extraction of domain-invariant embeddings.

Maximum Mean Discrepancy (MMD). The method proposed by(Long et al., 2015) focuses on minimizing the *Maximum Mean Discrepancy* (MMD) between source and target distributions. Given data from source and target distributions, $x^s \in \mathcal{S}$ and $x^t \in \mathcal{T}$ respectively, MMD is defined as:

$$MMD^2(\mathcal{S},\mathcal{T}) = \sup_{\|\phi\|_{\mathcal{H}}\leq 1} \left\|\mathbb{E}_{x^s\sim\mathcal{S}}[\phi(x^s)] - \mathbb{E}_{x^t\sim\mathcal{T}}[\phi(x^t)]\right\|^2_{\mathcal{H}} \tag{5.5}$$

where ϕ is a mapping function that belongs to a reproducing kernel Hilbert space $\mathcal{H}$. This approach encourages the final layers of the network to generate features that are invariant across domains.

Adaptive Feature Norm (AFN). (Xu et al., 2019c) identified that a key challenge in classifying target domain data is the tendency for target vectors to have smaller feature norms compared to those of the source domain. To address this problem, the authors proposed aligning the expected L_2-norms of the deep embeddings from both the source and target domains. More formally, defining $h(x) = \|x\|_2$, L_2-norm convergence is enforced by minimizing the following *Maximum Mean Feature Norm Discrepancy* (MMFND):

$$MMFND(\mathcal{S},\mathcal{T}) = \sup_{h\in\mathcal{H}} \frac{1}{n_s} \sum_{x_i^s\in\mathcal{S}} h(x_i^s) - \frac{1}{n_t} \sum_{x_i^t\in\mathcal{T}} h(x_i^t) \tag{5.6}$$

where $\mathcal{H}$ is the set of all functions composed by the L_2-norm operator, x_i^s and x_i^t are the i-th samples from the source and target domains respectively, n_s and n_t represent the total number of source and target samples in the source and target sets $\mathcal{S}$ and $\mathcal{T}$.

Rotation (ROT). Xu *et al.* (Jiaolong et al., 2019) introduced a novel approach to UDA that incorporates a self-supervised task involving geometric image transformations. This auxiliary task, which is solved in conjunction with the primary task, involves predicting the absolute rotation of images from both the source and target domains. The rotations are selected randomly from the set $\Theta = 0°, 90°, 180°, 270°$. This method assists the embedding model in better generalising across domains by leveraging the inherent structure of the images. Building on this, (Loghmani et al., 2020) expanded this concept to multi-modal images by designing a task where the network predicts the *relative rotation* between two modalities of the same input sample, such as an RGB image and its corresponding depth image. This extension aims to further enhance the model's ability to learn domain-invariant features by exploiting the relationship between different modalities of the same scene.

Entropy minimization (ENT). Entropy minimization (Grandvalet and Bengio, 2004) is a widely used technique to perform UDA. It consists of representing the uncertainty on the target domain through a functional that acts as a regularization term of the classification loss, which is referred to as *entropy loss*. More formally, the following loss is minimized during training:

$$\mathcal{L}_{\text{ENT}} = \mathcal{L}_y(\theta_f,\theta_y) - \frac{1}{|\mathcal{T}|} \sum_{x_t\in\mathcal{T}} F(x_t;\theta_f)\cdot \log G_y(F(x_t;\theta_f)), \tag{5.7}$$

where F and G_y are respectively the feature extractor and the label predictor, and x_t is a sample from the target distribution. Adding this functional as a regularization term of the classification loss helps soften the domain shift effects between source and target distributions.

Event-Representations

In this work, we focus on grid-like event representations, which entail converting a stream of asynchronous events into a volume $\mathcal{R}_\mathcal{E} \in \mathbb{R}^{H \times W \times F}$ with F features. We provide a summary of these representations below and refer the reader to Section 2.3.2 for a more comprehensive explanation.

Voxel Grids. This representation, also referred to as an event volume (Zhu et al., 2019a), divides time into a fixed number B of bins and aggregates events at their corresponding pixel locations by interpolating polarity values over time. The outcome is a B-channel representation where the contribution of each event is weighted based on the time of its occurrence within the temporal bin.

HATS. The Histograms of Oriented Time Surfaces (HATS) representation (Sironi et al., 2018a) is a two-channel approach that combines hand-crafted features with a mechanism resistant to noise. The event stream is partitioned into a grid of non-overlapping memory cells, each extracting local 2D surfaces from the vicinity of each event using an exponential kernel. These surfaces are then compiled into histograms, one for each polarity, and are organized based on the location of their originating cells. This method results in a loss of temporal resolution as the entire span of time is compressed into a single frame, thus not preserving the temporal detail.

EST. The Event Spike Tensor (EST) representation (Gehrig et al., 2019a) is end-to-end trainable. It operates similarly to a voxel grid, but with the distinction that timestamps serve as pixel features and the kernel function used to weigh the contribution of events is learned through a multi-layer perceptron network. Events are categorised by polarity to derive a two-channel representation from each temporal bin.

MatrixLSTM. MatrixLSTM (Cannici et al., 2020b) is similar to EST, with the primary distinction being that pixel features are computed using a matrix of LSTM cells (Hochreiter and Schmidhuber, 1997) with shared parameters. Each cell processes the time-ordered sequence of events produced by its corresponding pixel, and the final output of the LSTM serves as the pixel feature. The feature dimensionality is customisable, and temporal bins may be employed optionally to yield multiple representations.

5.3.3 N-ROD: a New Event-Based Dataset for Object Recognition

Collecting precisely annotated data presents a significant challenge, even when using standard vision devices. A common approach in the literature to circumvent this issue is the use of synthetic data generation, which offers the advantage of readily available, exact annotations. However, the disparity between synthetic training data and real testing data, known as the *synth-to-real* domain shift, significantly impacts the performance of the final model on real data. Domain adaptation techniques have proven to be an effective solution to this problem, as highlighted in several studies (Bousmalis et al., 2017; Sankaranarayanan et al., 2018; Vu et al., 2019). Yet, the specific effect of the *synth-to-real* shift on event data remains an underexplored area, largely due to the absence of suitable datasets for such analyses.

To fill this gap, we extend the widely-used RGB-D Object Dataset (ROD) (Lai et al., 2011) for object recognition with an event data counterpart. The original ROD dataset includes RGB and depth modalities captured with real sensors and has been recently augmented with synthetic samples (Loghmani et al., 2020). Building upon this, we introduce event data to both ROD and its synthetic variant, SynROD, facilitating *synth-to-real* studies for the event modality. This enhancement results in the creation of a new neuromorphic dataset, which we name N-ROD, offering comprehensive data for investigating the impact of *synth-to-real* shifts on event-based vision systems. Some examples from the proposed N-ROD dataset are shown in Figure 5.10.

Dataset

We propose an extension of the popular RGB-D Object Dataset (ROD) (Lai et al., 2011) for object recognition. ROD comprises 41,877 samples of 300 everyday objects organized into 51 categories, captured by an RGB-D camera. ROD is augmented with SynROD (Loghmani et al., 2020), its recent synthetic counterpart created to examine the synth-to-real domain shift in multi-modal contexts, such as RGB images and depth. SynROD includes photorealistic renderings of 3D models from the same categories as ROD, produced under natural lighting conditions. We enhance both versions of the dataset by incorporating real event recordings obtained from ROD samples, as well as simulated events derived from the synthetic images of SynROD. The augmented dataset thus created represents the first to facilitate a *synth-to-real* analysis on event data.

Recording Setup. We replicate the setting in (Orchard et al., 2015a) for converting RGB images to event-based recordings. A Prophesee's HVGA Gen3 (CD+EM) (Gallego et al., 2020b) Asynchronous Time Based Image Sensor (ATIS), configured with default

(h) Synthetic (k) Real

Fig. 5.10 **N-ROD examples.** Synthetic (left) and real (right) samples from the N-ROD dataset. Depth images are colorised with surface normal encoding and event sequences are represented using voxelgrid (Zhu et al., 2019a).

bias settings and equipped with a Computar M0814-MP2 8mm lens, is placed on a pan-tilt mechanism and positioned approximately 23 centimeters away from an LCD monitor. We used a 2560×1440 76Hz IPS monitor with a 4ms minimum response time (Lenovo^{TM} ThinkVision P27h-10), setting its brightness and contrast to their highest values as in (Hu et al., 2016). The pan-tilt mechanism[1], similar to the one used in (Orchard et al., 2015a), consists of two Dynamixel MX-28 servo motors interconnected, controlled by an ArbotiX-M Robocontroller board via serial communication.

Objects from the ROD dataset are presented in crops of variable size and aspect ratio. To process the samples, padding is applied, replicating the border on the shorter side of the image to ensure squared samples, regardless of the original resolution. Still images from the original ROD dataset are displayed in a loop, and each sample is recorded while performing the same saccadic motion pattern described in (Orchard et al., 2015a) (i.e., three saccadic motions of 100ms each, forming a triangular pattern). A waiting period of 300ms is added after transitioning to the next image to guarantee that the image is correctly updated on the monitor and that the event camera has stabilised after detecting the visual changes induced by the image switch. A 256×256 region of interest is designated on the event camera to limit recorded events to a squared resolution, mirroring the ROD RGB images. Grayscale images from exposure measurement (EM) events are utilized to adjust the size of displayed images to the camera's field of view before recording.

To simulate data in the source domain, we follow the procedure outlined by(Gehrig et al., 2020), utilising the ESIM simulator (Rebecq et al., 2018) to generate events. We replicate

[1] https://trossenrobotics.com/widowx-MX-28-pan-tilt

the same settings employed for recording real samples, by projecting synthetic images onto a plane and moving the virtual event camera through the usual saccadic motion.

5.3.4 Experiments

We conduct experiments on the object classification task. We utilize the N-Caltech101 dataset to compare with state-of-the-art approaches under the E-Sim-to-Real shift. We then employ the proposed N-ROD dataset to evaluate the DA4E framework under both the RGBE-Synth-to-Real and RGB-Synth-to-Real settings. The proposed DA4E is assessed using the UDA methods described in Section 5.3.2, and employing the event representations detailed in the same section. In the following sections, we detail the datasets used and then discuss the experimental validation conducted.

Datasets

Apart from the N-ROD dataset already discussed in the previous section, we conduct single-modal experiments on N-Caltech101 (Orchard et al., 2015a).

N-Caltech101. The Neuromorphic Caltech101 (N-Caltech101) dataset (Orchard et al., 2015a) represents an event-based conversion of the well-known image dataset Caltech-101 (Fei-Fei et al., 2006). Samples from N-Caltech101 were generated by capturing the original RGB images with a real event-based camera, which was moved in front of a stationary monitor displaying the images. A recent extension to N-Caltech101 has been introduced in (Gehrig et al., 2020), wherein a simulated replica of the dataset was created using the ESIM simulator (Rebecq et al., 2018). This process involved re-creating the same setup utilized for capturing the real samples. Following the approach in (Gehrig et al., 2020), we use these recordings as simulated source data and those from N-Caltech101 as the real target samples. We use the train and test splits provided in the EST (Gehrig et al., 2019b) official codebase, and evaluate the proposed approach by computing the top-1 accuracy on the test set of the target real domain, as in (Gehrig et al., 2020).

Implementation details

We implement the proposed method within the PyTorch autodiff framework, employing a ResNet34 (He et al., 2016) as the feature extractor $\mathcal{F}$ in N-Caltech101 experiments, and a ResNet18 (He et al., 2016) for N-ROD experiments, both pre-trained on ImageNet. To ensure

a fair comparison, we adopt the same network configurations as in (Loghmani et al., 2020) for both the object recognition classifier $\mathcal{G}$ and the network utilized in the pretext rotation task. The proposed multi-view approach is compared against a baseline with identical architecture, pre-trained on ImageNet, but wherein event representations are directly inputted as a singular multi-channel tensor without view grouping. Here, the first convolutional layer is substituted with a newly, randomly initialised convolution to match the input channels' number, and the multi-view pooling stage is omitted. Event representations and RGB images processed through the main backbone $\mathcal{F}$ are preprocessed and augmented during training as per the procedure in (Loghmani et al., 2020). Input images are normalized using the same mean and variance as for ImageNet pre-training, while event representations are kept unnormalized, as this yielded better results. We utilize 9 bins for both voxel grids and EST representations, resulting in 3 and 6 views respectively, given that the latter generates 2 channels per bin. The output channels' number can be tailored in MatrixLSTM, hence we configure the layer to directly produce 3-channel output representations and set the bin count to 3, as this setup showed optimal performance. Given that HATS solely offers 2 channels, without default temporal frame splitting into bins, the proposed multi-view approach is inapplicable. All network configurations are trained using SGD as the optimiser, with a batch size of 32 and 64 for N-Caltech101 and N-ROD experiments, respectively, and a weight decay of 0.003. The weights of the DA losses for each event representation and DA method are fine-tuned, reporting only the accuracy scores of the best configurations, averaged over 3 runs with different random seeds.

E-Sim-to-Real results

We initially evaluate the effectiveness of the UDA algorithms in mitigating the domain shift under the E-Sim-to-Real scenario using N-Caltech101. In Table 5.6, we present the performance of GRL (Ganin and Lempitsky, 2015b), MMD (Long et al., 2015), Rotation (Jiaolong et al., 2019), AFN (Xu et al., 2019c), and Entropy (Grandvalet and Bengio, 2004) compared to the baseline Source Only, which is the network training on labeled source data only (*Sim*) and testing directly on unlabeled target data (*Real*) without any adaptation strategy. We use the performance achieved by training on real training data and testing on it in a supervised manner (*Supervised*) as the upper bound. For each method, we report results both with (*MV-DA4E*) and without (*DA4E*) the proposed multi-view approach. The impact of UDA strategies on two non-learnable event representations (*VoxelGrid* and *HATS*), and two learnable ones (*EST* and *MatrixLSTM*) is considered.

N-Caltech101 (Sim $\Longrightarrow$ Real)					
Method		Voxel Grid	HATS	EST	Matrix LSTM
Source Only	*baseline*	80.99	58.32	80.08	82.21
	MV-baseline	84.59	-	83.07	84.89
GRL (Ganin and Lempitsky, 2015b)	DA4E	83.08	65.38	83.38	82.94
	MV-DA4E	86.77	-	84.03	85.75
MMD (Long et al., 2015)	DA4E	86.37	69.86	83.61	84.04
	MV-DA4E	88.23	-	85.36	**88.05**
Rotation (Jiaolong et al., 2019)	DA4E	79.13	61.52	80.69	83.57
	MV-DA4E	86.63	-	84.49	85.7
AFN (Xu et al., 2019c)	DA4E	84.49	**69.96**	83.59	85.0
	MV-DA4E	88.3	-	85.92	87.59
Entropy (Grandvalet and Bengio, 2004)	DA4E	87.0	65.58	85.54	85.97
	MV-DA4E	**89.24**	-	**86.06**	86.09
Supervised	*RealEvent*	88.13	76.45	88.17	87.65
	MV-RealEvent	90.09	-	89.25	90.35

Table 5.6 **Results on N-Caltech101.** Target Top-1 Test Accuracy (%) of UDA methods on N-Caltech101. Bold: representation's highest result.

UDA results. From the results in Table 5.6 it can be noted that, for all event representations, in almost all cases the UDA methods outperform the baseline Source Only, exceeding it by up to 6% on VoxelGrid, 11% on HATS, 6% on EST, and 4% on MatrixLSTM. There is a single instance where Rotation is on par with the Source Only, which is the case for VoxelGrid without the multi-view approach. This may be because the principal advantage of Rotation is to push the network to focus on the geometric aspects of the input through solving the transformation. Given that event data inherently encodes geometric information (e.g., direction of movement), Rotation might, in some scenarios, be potentially unhelpful. Indeed, the network could learn to identify a trivial solution (shortcut) for solving the pretext task (Noroozi and Favaro, 2016), for example by analysing the direction of movement across edges. Interestingly, it can be observed that not all representations are equally affected by the domain shift. For example, HATS is the representation most affected by the Sim-to-Real shift, with performance decreasing by up to 16% when testing directly on the target domain (Source Only) rather than on the source (Supervised). Intuitively, the reason lies within the representation itself. In fact, when events are represented using HATS, the temporal resolution is lost (see Section 5.3.2), potentially leading to a degradation in performance when testing on data from a different distribution.

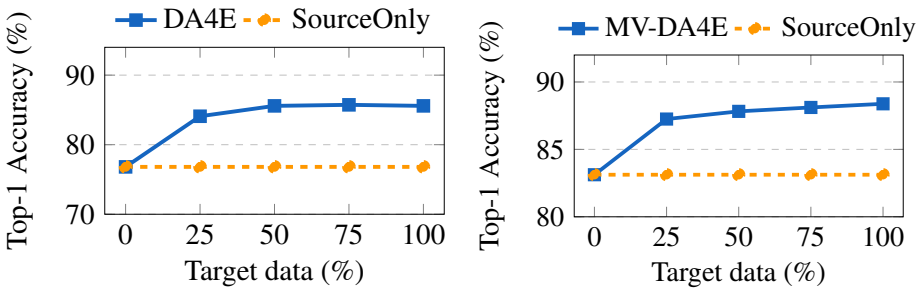
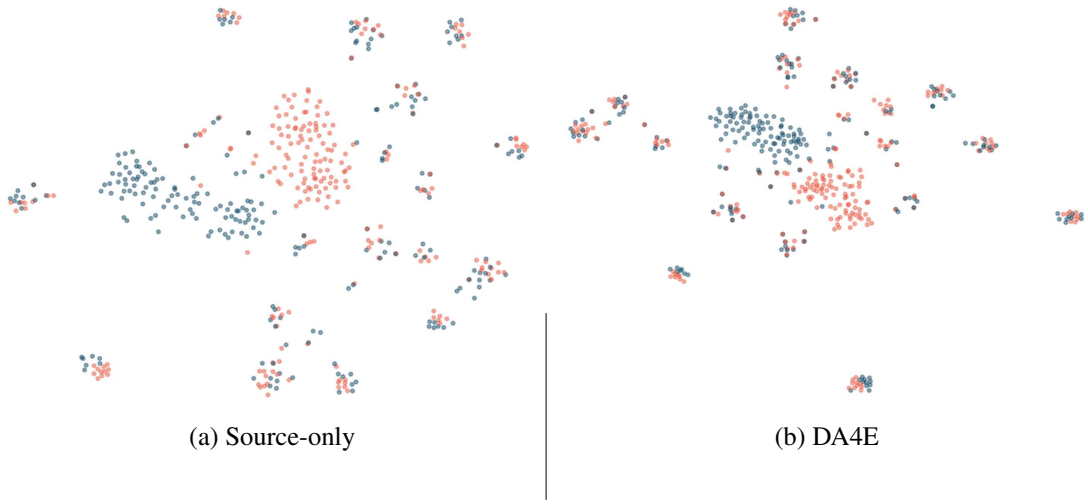


Fig. 5.11 **Ablation on percentage (%) of target.** Difference in terms of performance based on percentage (%) of target data used during training, obtained with constant threshold $C = 0.06$.

In Figure 5.11, we demonstrate the scalability of our approach when access to target data is limited, by illustrating how the performance of the proposed methods varies when only a percentage of target data is available during training (25%, 50%, 75%). It is noticeable that an improvement of up to 4% over the Source Only baseline (0% of training target data) is assured, even when a very small percentage of target samples is available. Qualitative results are presented in Figure 5.12, where we provide a t-SNE visualization of the source and target samples, both when adapting the two domains and when not adapting them. We also computed the Gradient-weighted Class Activation Mapping (Grad-CAM (Selvaraju et al., 2017)) on several N-Caltech101 samples, which visualizes regions in the input event representation upon which the network most significantly focuses for prediction. As illustrated in Figure 5.13, when trained with the proposed MV-DA4E approach, these regions are the most discriminative for classifying the object.

MV-DA4E. Table 5.6 demonstrates that applying the multi-view approach *MV-DA4E* significantly enhances performance compared to the *DA4E* configuration across all experiments, independently by representations and DA strategies utilized. These results validate the effectiveness of the proposed method, corroborating the assertions made in Section 5.3.2. Interestingly, *MV-DA4E* not only facilitates improvement in the cross-domain scenario (Sim-to-Real) but also within the intra-domain (Supervised) context. Consequently, we show that this multi-view approach could serve as a universally applicable strategy for managing event representations, regardless of the specific task being addressed.

Comparison with approaches acting on the threshold C. Several methods in the literature, such as (Gehrig et al., 2020; Stoffregen et al., 2020), address the *Sim-to-Real* challenge by primarily manipulating the threshold value C utilized by the simulator for data generation. Since our approach uses a fixed threshold, it raises the question of whether our results



(a) Source-only (b) DA4E

Fig. 5.12 **t-SNE visualization.** t-SNE visualization of N-Caltech (Orchard et al., 2015a) features from the last hidden layer of the main classifier. Red dots: source samples; blue dots: target samples. When adapting the two domains with the proposed DA4E (b), the two distributions align much better compared to the non-adapted case (a).

are attributed to an optimal selection of C or if they are a consequence of our decision to promote adaptation by focusing on feature-level modifications. To answer this question, we conducted experiments with different UDA methods using the voxel grid representation and three different settings for C, namely $C = 0.06$ (the initial value used to examine the domain shift in (Gehrig et al., 2020)), $C = 0.15$ (determined following (Stoffregen et al., 2020)), and $C \sim \mathcal{U}(0.05, 0.5)$ (as suggested in (Gehrig et al., 2020)). The baselines include methods based solely on C (specifically, $C = 0.15$ replicates the conditions in (Stoffregen et al., 2020) and $C \sim \mathcal{U}$ those in (Gehrig et al., 2020)). Results presented in Table 5.7 reveal that: (i) our methodology consistently and significantly outperforms the baselines across every tested value of C, underscoring the advantage of addressing DA at the feature level; (ii) the multi-view strategy benefits from UDA techniques in every scenario; and (iii) even the methods based solely on C gain from adopting a multi-view approach, as it markedly mitigates their sensitivity to variations in C.

RGBE and RGB Synth-to-Real results

In robotics, Domain Adaptation (DA) leverages automatically generated synthetic data with "free" annotations to enhance predictions on real data and offset the absence of extensive datasets. Nevertheless, differences between synthetic training data and real test data, commonly referred to as the *synth-to-real* domain shift, severely undermine the final model's

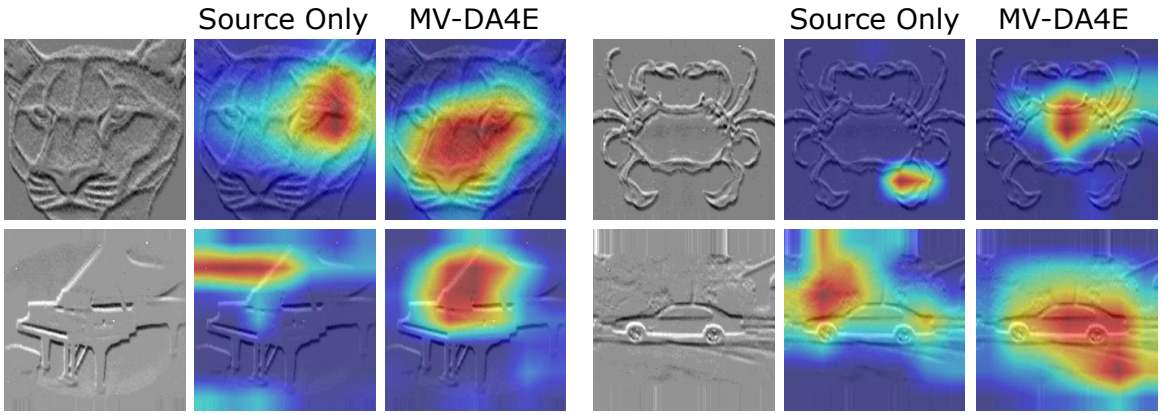


Fig. 5.13 **Grad-CAM (Selvaraju et al., 2017) visualizations.** Grad-CAM (Selvaraju et al., 2017) visualizations on several real N-Caltech101 samples. In each triplet we show the input event representations (voxel grid (Zhu et al., 2019a)), the activation maps when the network is trained on simulated data only, and those obtained by training with MV-DA4E.

performance on the actual data. It has been shown that leveraging the complementary nature of multi-modal inputs can enhance adaptation performance in cross-domain scenarios (Loghmani et al., 2020). To explore the efficacy of event data derived from RGB images and their applicability in real-world settings, we investigate the performance of the event modality (both as a standalone and when combined with RGB in a multi-modal RGB+Event setup) on the N-ROD dataset. This analysis aims to evaluate the advantages of the event modality compared to traditional ones, such as RGB and depth.

For this analysis, we selected the VoxelGrid representation, which demonstrated superior cross-domain performance, and the multi-view approach *MV-DA4E*, which proved to be highly effective in all conducted experiments on event data (Table 5.6).

RGBE-Synth-to-real. Table 5.8 presents the results in a RGBE-Synth-to-Real scenario, where events simulated on synthetic data ($ESIM(RGB_{synth})$) serves as the source and events recorded with a neuromorphic camera from real RGB images ($EvCamera(RGB_{real})$) is the target.

The findings indicate that the event modality benefits the most from UDA, with a 20.6% improvement over the Source Only approach, while RGB and depth modalities show smaller gains of 9.9% and 14.4%, respectively. The event modality's focus on geometric components and object shapes—unlike RGB's texture bias—makes UDA techniques particularly effective for events. This is because shape information is inherently more robust in transitioning from

N-Caltech101				
Baselines		C=0.06	C=0.15	C~ $\mathcal{U}$
Source only	*baseline*	76.81	80.99	82.29
	MV-baseline	83.12	84.59	84.93
Our approach	w/ C values:	C=0.06	C=0.15	C~ $\mathcal{U}$
GRL (Ganin and Lempitsky, 2015b)	DA4E	80.89	83.08	81.91
	MV-DA4E	84.93	86.77	86.45
MMD (Long et al., 2015)	DA4E	83.84	86.37	84.38
	MV-DA4E	86.94	88.23	87.31
ROT (Jiaolong et al., 2019)	DA4E	80.05	79.13	80.36
	MV-DA4E	86.31	86.63	87.08
AFN (Xu et al., 2019c)	DA4E	84.38	84.49	84.3
	MV-DA4E	87.71	88.3	88.17
Entropy (Grandvalet and Bengio, 2004)	DA4E	85.26	87.0	85.16
	MV-DA4E	**88.38**	**89.24**	**88.61**

Table 5.7 **Comparison with approaches acting on the threshold C.** Target Top-1 Test Accuracy (%) of UDA methods w.r.t. to methods that act on the contrast threshold C.

synthetic to real domains, facilitating alignment more so than the information encoded in RGB images.

The literature acknowledges the advantage of leveraging the complementary nature of different input modalities, such as RGB and depth, to enhance adaptation performance in cross-domain settings. With multi-modal RGB-E (RGB and Event) analysis still uncharted in research, we introduce an initial approach to this challenge, inspired by strategies used for RGB-D (RGB and Depth) data. Results underscore the efficacy of DA strategies across both single and multi-modal settings, with all methods showing consistent improvements over the Source Only baseline. Notably, the "Rotation" method, when applied to each modality individually, yields the least performance gain among the methods tested. However, when adapted to the RGB-E context through "Relative Rotation" between modalities, it surprisingly outperforms other UDA techniques. This highlights the significance of exploiting the complementarity between modalities, even within the realm of event data, suggesting a promising direction for future research in developing networks that efficiently integrate these two modalities.

RGB-Synth-to-real. Employing simulations on one side and actual event data on the other leads to the introduction of a Sim-to-Real discrepancy, as explored in (Gehrig et al., 2020;

$ESIM(RGB_{synth}) \implies EvCamera(RGB_{real})$					
	Single-modal			Multi-modal	
Method	RGB	Depth	**Event**	RGB+D	**RGB+E**
Source Only	52.13	7.56	21.78	47.70	50.78
GRL (Ganin and Lempitsky, 2015b)	57.12	26.11	33.09	59.51	57.15
MMD (Long et al., 2015)	63.68	29.34	42.05	62.57	61.78
Rot (Jiaolong et al., 2019)(Loghmani et al., 2020)	63.21	6.70	31.26	66.68	68.54
AFN (Xu et al., 2019b)	64.63	30.72	55.12	62.40	64.04
Entropy (Grandvalet and Bengio, 2004)	61.53	16.79	50.14	63.12	64.08
Avg	62.03 ▲+9.9	21.93 ▲+14.4	42.33 ▲+20.6	62.86 ▲+15.2	**63.12** ▲+12.3

Table 5.8 **Top-1 accuracy (%) of UDA methods on RGBE-Synth-to-Real shift. Bold:** highest mean result, underline: highest single- and multi-modal results. ▲ indicates the improvement of the avg of UDA methods over the baseline Source Only.

$ESIM(RGB_{synth}) \implies (ESIMvsEvCamera)(RGB_{real})$								
Source	Target	Source Only	GRL	MMD	Rot	AFN	Entropy	Avg
Sim	Real	21.78	33.09	42.05	31.26	55.12	50.14	42.33
Sim	Sim	40.47	44.52	48.29	42.98	53.50	49.29	**47.68**

Table 5.9 ***sim-to-real* and *sim-to-sim* scenarios.** Top-1 accuracy (%) on events, in two different scenarios: *sim-to-real* and *sim-to-sim*. In **bold** the highest mean result.

Stoffregen et al., 2020). To assess the impact of this additional domain shift on performance, we compare our findings with outcomes derived from simulating events from actual (target) images ($ESIM(RGB_{real})$), thus creating a scenario where the Sim-to-Real gap is not present.

For this purpose, Table 5.9 presents a comparison of our single-modal results, where target events are captured using a neuromorphic camera (Source: Sim, Target: Real), against those generated entirely through simulation (Source: Sim, Target: Sim).

By observing the Source Only results, we observe a performance drop by up to 20%, clearly showing the Sim-to-Real impact and underscoring the necessity for methodologies to bridge the simulated and real-world data gap. Once more, our strategy demonstrates the value of applying UDA techniques within the realm of event data, markedly enhancing performance and diminishing the Sim-to-Real gap to a mere 5%.

5.4 Conclusion

In this chapter, we introduced N-EPIC-Kitchens, the first event-based egocentric action recognition dataset. By leveraging the variety of data modes at our disposal, we conducted an in-depth comparative analysis, the results of which underscore the significance of motion information in the context of action recognition. Based on these findings, we proposed and evaluated two innovative approaches tailored for event data (E^2(GO) and E^2(GO)MO) that, by highlighting motion information, yielded competitive results compared to the computationally expensive optical flow modality. Our extensive experiments shed light on the robustness of event data and their suitability for an online action recognition scenario, encouraging the community to delve further into this area.

Introducing event data in egocentric action recognition for the first time, we aim to provide a direct comparison with established benchmarks in the literature (Damen et al., 2018, 2022; Fathi et al., 2012b), positioning the event modality competitively against well-established modalities. This objective motivated us to simulate event data rather than creating a new first-person dataset from scratch. To justify our choice, we then proposed an alternative approach to a very recent research problem: how to bridge the Sim-to-Real gap for event cameras that arises from event generation. By viewing the problem from a new perspective—namely, the domain shift—we demonstrated that Unsupervised Domain Adaptation (UDA) techniques operating at the feature level are an effective way to address this issue, compared to previous work that focused on the input level. Additionally, we introduced a multi-view approach for handling event representations, which outperforms existing methods and proves to work well in conjunction with other UDA strategies.

We demonstrate that despite its high computational and temporal costs, the TV-L1 optical flow still shows superior performance, particularly an exceptional resilience to domain changes. We primarily attribute this to the algorithm's ability to partially filter out camera motion, yielding cleaner motion data compared to raw events. Future work could explore using motion compensation techniques commonly employed with event data (Stoffregen et al., 2019) to eliminate redundant background noise. Moreover, building on the promising results of our work, we plan to further investigate the use of real event streams in this context to confirm the insights gained so far with a real camera.

Chapter 6

Egocentric Video Understanding using 3D

In the previous chapters, we addressed issues related to cross-domain challenges or the expensive computation of traditional modalities. In this chapter, we introduce the use of 3D information to address the limitations of the narrow field of view of egocentric devices. The egocentric viewpoint is inherently limited, primarily due to the recording device's proximity to the location where interactions occur. This means that at any moment, the camera captures only a small portion of the broader scene, significantly restricting our understanding of the scene in its entirety. This challenge is further amplified by the dynamic nature of human interaction with their surroundings: as the individual wearing the camera handles objects, these items frequently move in and out of the camera's field of view. This frequent movement not only complicates understanding the events within the scene but also challenges tracking objects once they have moved beyond the immediate field of vision, a concept referred to as *object permanence*.

We propose to combine 2D frame-based information captured by the camera with 3D information about the scene and the locations of objects within it. This integration enriches egocentric vision with the capability to remember the locations of objects even when they are no longer visible in the egocentric video stream. This capability, commonly referred to in humans as "spatial cognition", forms the basis for our introduction of the task "Out of Sight, Not Out of Mind" (OSNOM)—maintaining knowledge of where objects are, even when they are absent from the video stream.

The work in this chapter can be found in the following article:

- Plizzari, C., Goel, S., Perrett, T., Chalk, J., Kanazawa, A., Damen, D. (2024). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. *Preprint.*
Online Resources: [Paper], [Project page]

6.1 Introduction

It's lunchtime, and the pan is on the stove. You bend down to pick up the chopping board from a lower cupboard and place it on the counter. Then, you retrieve a knife from the cutlery drawer. You use the chopping board and knife to slide the chopped food into the pan before tossing both into the sink. Afterwards, you grab a clean plate from the drainer to serve the food. As you move around the kitchen, you are aware of where these objects are, even if they are currently out of view.

Spatial cognition allows humans to construct a mental map of their surroundings, which includes "memories of objects once perceived as we moved about" (Downs and Stea, 1973). Importantly, spatial cognition posits that these objects exist independently of human attention and continue to exist on the cognitive map even after the observer has left the vicinity (Burgess, 2006; Committeri et al., 2004; Moore and Meltzoff, 2004; Zewald and Jacobs, 2022). Spatial cognition is an innate ability that is crucial for human survival; it enables individuals to "acquire and use knowledge about their environment to determine their location, how to obtain resources, and how to find their way back home (Waller and Nadel, 2013)."

In this chapter, we introduce the task "Out of Sight, Not Out of Mind" (OSNOM) – maintaining the knowledge of where *all objects* are located, even as they move and when they are absent from the egocentric video stream. Egocentric views facilitate detailed observations of object interactions, such as looking into a fridge or oven, and identifying items removed from a drainer. Nonetheless, objects frequently exit the camera's field of view due to the movements of the person wearing the camera. We focus on this challenging set of active objects that move within the video sequence. Figure 6.1 illustrates the OSNOM task, where the 3D locations of objects and their movements are tracked throughout the video, irrespective of the objects' visibility. To address the OSNOM challenge, we introduce a method that *lifts* 2D observations into a 3D world coordinate frame. This is achieved by reconstructing the scene mesh and projecting 2D observations using their depth relative to the camera and estimated surfaces. We then *match* these transformed observations based on appearance and location over time to establish consistent object tracks, and maintain awareness of objects even when they are not visible. This *lift, match, and keep (LMK)* approach facilitates egocentric spatio-temporal understanding by combining 2D partial view with 3D information about object locations from egocentric videos.

In summary, the contributions of this chapter are the following:

- We introduce Lift, Match and Keep (LMK) to address the OSNOM challenge. That consists in lifting objects in 3D, matching their 3D location, and use it to keep them

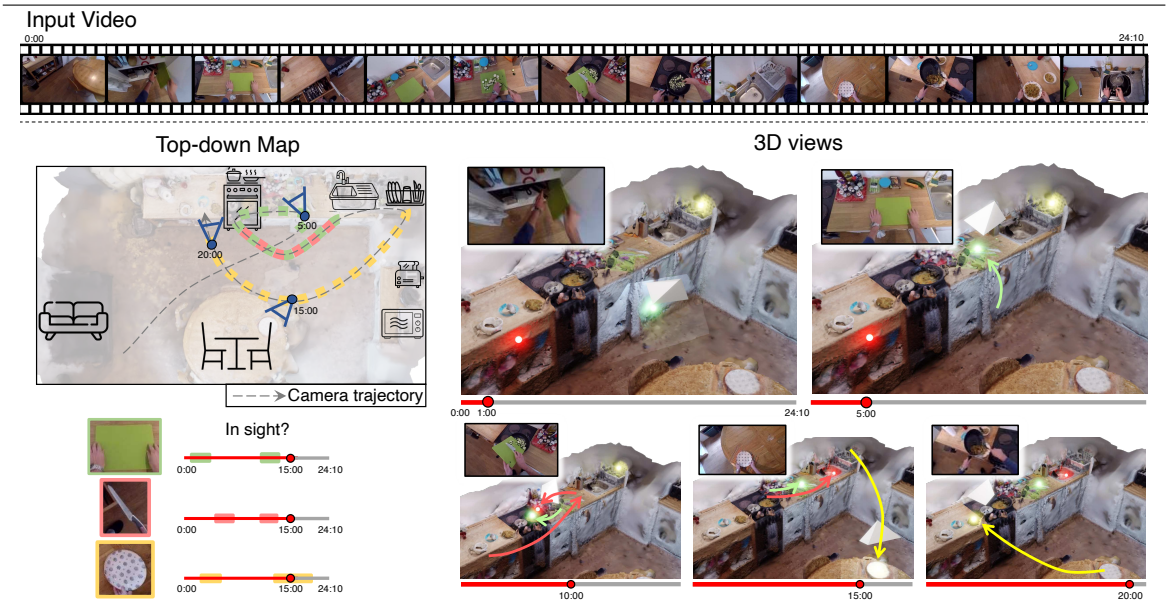


Fig. 6.1 **Spatial Cognition.** From an egocentric video (top), we introduce the task "Out of Sight, Not Out of Mind", which entails tracking the 3D locations of all active objects, visible or not. We present a 24-minute video to demonstrate how this task aids in tracking three active objects throughout the video within a global coordinate system. This includes a top-down view featuring camera movement (top left), the identification of moments when objects are visible (bottom left), and their trajectories from a side view across five different frames (right). Neon balls indicate the 3D locations of these objects over time, alongside the camera (represented as a white prism), the corresponding frame (inset), and changes in object locations (colored arrows). The chopping board is retrieved from a lower cupboard at 1:00 and is in hand by 05:00. The knife is taken from the drawer shortly after 05:00, used by 10:00, and then discarded into the sink before 15:00. The plate moves from the drainer to the table at 15:00, and then back to the counter by 20:00.

tracked over time using both 2D egocentric videos and locations of objects in 3D (Section 6.3);

- We evaluate our approach using 100 videos from the EPIC-KITCHENS dataset (Damen et al., 2022), assessing past and future 3D location estimations across multiple time scales. Our results show that objects are out of view in approximately 85% of the frames on average. With our LMK approach, we are able to accurately position 64% of the objects after one minute, 48% after five minutes, and 37% after ten minutes. Results of the LMK on the OSNOM task are presented in Section 6.4.1.

6.2 Background

Traditionally, egocentric vision has focused on tasks relying solely on the recorded video stream, i.e., within the camera's field of view. These tasks range from understanding actions,

objects, and interactions over short, and more recently, longer timescales (Damen et al., 2022; Darkhalil et al., 2022; Grauman et al., 2022; Tang et al., 2024). Even when addressing future predictions (e.g., action anticipation in (Girdhar and Grauman, 2021)), memory (e.g., episodic memory in (Grauman et al., 2022)), or object tracking (Tang et al., 2024), these approaches scan the video stream to determine when an object is in sight. The seminal work Ego-Topo (Nagarajan et al., 2020) builds a 2D affordance graph of the environment, relating actions to automatically discovered hotspots. The motivation for capturing the relative location of an object to the camera wearer was further explored in the EgoEnv paper (Nagarajan et al., 2024), where pre-training was conducted on 3D simulated environments. This shows that environmentally-informed representations can enhance performance on downstream tasks such as episodic memory. A number of tasks have recently been proposed that require **3D understanding in egocentric vision**, such as jointly recognizing and localizing actions in a 3D map (Liu et al., 2022). A task related to ours is Visual Query Localization in 3D (VQ3D) (Grauman et al., 2022). A recent approach, EgoLoc (Mai et al., 2022), searches for the last frame in which the query object appears through 2D detection and proposes an improved pipeline to determine the 3D location of this single object. In contrast to (Mai et al., 2022), which aims to determine the 3D location of one in-view object at a single moment, OSNOM seeks to ascertain the 3D locations of multiple objects over time, even when they are in-hand, moving, occluded, or out of the camera's view.

3D Egocentric Datasets are now increasingly available, as evidenced by sources such as ODIN (Ravi et al., 2023), Ego4D (Grauman et al., 2022), Aria Digital Twin (Pan et al., 2023), and EPIC-Fields (Darkhalil et al., 2022). We refer to Section 2.1.3 for a broader overview on those datasets. EPIC-Fields (Darkhalil et al., 2022) offers a comprehensive pipeline for extracting point clouds and dense camera poses from egocentric videos. This pipeline facilitates camera estimates for videos from the EPIC-KITCHENS dataset (Damen et al., 2022) across 45 kitchens and pairs them with dense active object masks from VISOR (Darkhalil et al., 2022). In this study, we employ the EPIC-Fields pipeline to localize cameras within the world coordinate frame and utilize VISOR masks to identify active objects.

Object tracking through occlusion has been extensively studied in 2D. Maintaining object permanence through heuristic methods, such as assuming constant velocity (Breitenstein et al., 2009), or through learning-based approaches (Shamsian et al., 2020; Tokmakov et al., 2021), facilitates the reassignment of tracks when occluded objects reappear (Huang and Essa, 2005). However, these works do not track objects outside of the camera's field of view, and the datasets specifically targeting occlusion are short-term, especially those with non-synthetic footage (e.g., the recent TCOW (Van Hoorick et al., 2023) has a maximum video length of 464 *frames*).

Autonomous driving maintains a map of the vehicle's surroundings (Wong et al., 2020), allowing it to track nearby vehicles, even when out of sight. While these systems maintain knowledge of surrounding objects through occlusion (Gilroy et al., 2019; Ren et al., 2021), tracks are deleted after a short time as the vehicle only needs to be aware of objects within its vicinity.

Human tracking has evolved from 2D (Bergmann et al., 2019; Meinhardt et al., 2022; Zhang et al., 2022c), to 3D (Rajasegaran et al., 2021), and further to 3D with motion models (Goel et al., 2023; Khurana et al., 2021; Rajasegaran et al., 2022), which predict the locations of occluded humans. Although these approaches utilize 3D for tracking, they typically do so within the camera coordinate frame. Recent studies have begun to explore the simultaneous reconstruction of camera motion and human pose within the 3D world coordinate frame (Kocabas et al., 2023; Ye et al., 2023; Yuan et al., 2022). Notably, (Sun et al., 2023) and (Ye et al., 2023) have applied this concept to the tracking of human subjects. (Khirodkar et al., 2023) introduces a benchmark for tracking humans from multiple ego- and exo-centric camera perspectives.

Our approach aligns with these advancements in human 3D tracking. We present the first egocentric vision effort that focuses on tracking objects within the world coordinate frame. Unlike humans, objects in egocentric videos do not move by themselves and are thus considered static when not being manipulated by the camera wearer. Conversely, these objects frequently move in and out of the camera's view and are often occluded or blurred. We expand upon the human tracking methodology (Rajasegaran et al., 2022), adapting it to track objects while leveraging camera localization within the environment. Differing from previous works, we maintain and assess the 3D world coordinates of objects even when they are out-of-view.

6.3 Method - Lift, Match and Keep (LMK)

Our method takes in input an untrimmed/unedited egocentric video, which we refer to as E, that has been recorded in an indoor environment. Our ultimate goal is to maintain continuous tracking of all objects of interest within the 3D world coordinate frame. By consistently capturing the locations of all objects—even when they are not visible in the camera frame—these 3D tracks address the challenge of "Out of Sight, Not Out of Mind" (OSNOM). We focus on the challenging set of objects the camera wearer interacts with–moving them from one place to another, often multiple times in the video–rather than the

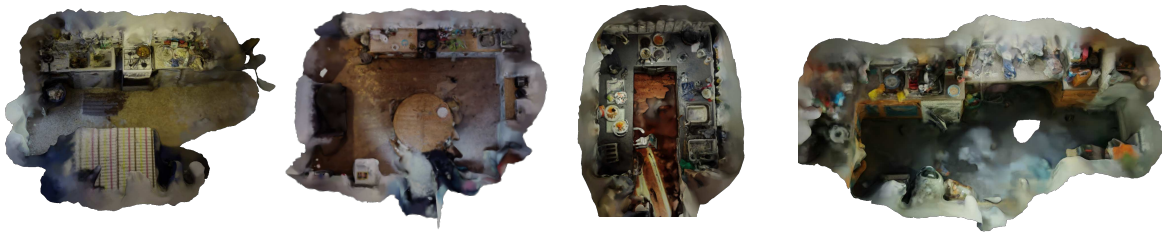


Fig. 6.2 **3D reconstruction of the scenes.** Example of 3D meshes of 4 different environments using Poisson surface reconstruction.

objects in the scene that remain in the same position throughout the entire video. We refer to these as *active objects*.

Our method takes as input observations of active objects $o_n = (f_n, m_n)$, where f_n indicates a frame, and m_n is a semantic-free 2D mask in that frame given in *image coordinates*. Throughout the entire video, the set of observations is $O = \{o_n : n = 1, ..., N\}$. Since these observations are not present for every object and in every frame, i.e., only objects in the camera frame have a corresponding observation, we refer to them as *partial* observations. Since each object may be associated to multiple observations, the number of observations N is significantly greater than the number of active objects. Due to the possibility of a frame having zero or more masks, N is also independent of the number of frames T.

We name our method Lift, Match and Keep (LMK). First, we *lift* 2D observations of objects to 3D world coordinates, *match* them over time, and *keep* objects in mind when they are out-of-sight.

In Section 6.3.1 we describe the process of lifting our 2D observations into the 3D world coordinate frame by reconstructing the global 3D representation of the static scene along with registered camera poses for each frame. Section 6.3.2 explains how we use 3D distances and visual appearance similarity to match these lifted observations in 3D across frames. We preserve 3D observed locations when the objects are out of sight, which is crucial for OSNOM. Section 6.3.3 describes how we can specify object properties explicitly in relation to the camera wearer and environment using information from LMK.

6.3.1 Lift: Lifting 2D Observations to 3D

3D Scene Representation. Given a single egocentric video stream as input, we use the pipeline described in (Tschernezki et al., 2024) to estimate camera poses and a sparse point cloud of the static scenes. By computing the homography over consecutive frames, we eliminate redundant frames, enabling Structure from Motion (SfM) pipelines like COLMAP (Schön-

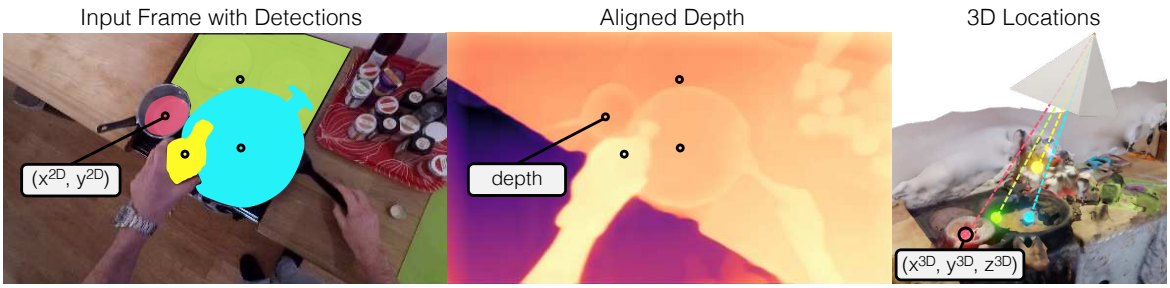


Fig. 6.3 **Lifting 2D observations to 3D.** An example of lifting multiple objects from a 2D image to 3D world coordinates, using masks, the camera pose, and a reconstructed mesh of the environment.

berger and Frahm, 2016) to handle long videos. The resulting subset of video frames contains enough visual overlap to register all frames to the SfM point cloud and estimate a camera pose C_t for every time t in the video. Note that this pipeline automatically estimates the intrinsic parameters of the camera.

This reconstruction focuses on estimating the static background of the scene. Objects in motion are treated as outliers during matching and are accordingly ignored in the reconstructions. Since the pipeline has no knowledge of surfaces, it generates a sparse point cloud that is unusable for 3D object positioning. Thus, we convert these point clouds to surface representations as follows.

We utilize a conventional Multi-View Stereopsis pipeline (Furukawa and Ponce, 2009; Schönberger et al., 2016) to generate scene geometry as a 3D mesh. This involves performing patch matching to establish dense correspondences between stereo image pairs, triangulating these correspondences to infer depth, and then amalgamating them into a dense 3D point cloud with surface normals. We then apply Poisson surface reconstruction (Kazhdan et al., 2006) to successfully derive a scene mesh $\mathcal{S}$ from the point cloud. Figure 6.2 showcases examples of these meshes.

Estimating 3D locations. For each frame f_n, we estimate the corresponding monocular depth using a very recent monocular depth estimation pipeline (Yang et al., 2024). The advantage of this approach lies in its capability to accurately estimate the positions of both static and dynamic objects, including those that are being held in-hand. However, this per-frame depth is temporally inconsistent across frames and not scaled with respect to real depth values in the 3D world. We use rendering techniques to estimate mesh's depth from a given camera viewpoint. We then apply a scale-shift transformation that minimizes the least squares error between the depth estimate through monocular depth estimation and the mesh's rendered depth. The results is an *aligned depth map*.

Given an observation $o_n = (f_n, m_n)$, we consider the centroid of the 2D mask m_n as the object's 2D location. Note that we represent each observation as a point in 3D following previous works (Grauman et al., 2022; Mai et al., 2022). We then take the depth value d_n in correspondence to the object's 2D location on the aligned depth map, and assign d_n to observation o_n.

Given the object's 2D location in frame f_n, depth relative to the camera d_n, and camera pose C_{f_n}, we project the observation to the 3D world coordinate as in the following:

$$[X_n, Y_n, Z_n]^T = C_{f_n} \begin{bmatrix} d_n K^{-1} [x_n, y_n, 1]^T \\ 1 \end{bmatrix} \tag{6.1}$$

where X_n, Y_n, Z_n represent the resulting 3D location of observation o_n, and K represents the camera's intrinsic parameters.

We denote this 3D location as $l_n \in \mathbb{R}^3$. The process of *lifting* to 3D is visualized in Figure 6.3. At this stage, these 3D observations are still partial and confined to individual frames.

Visual features. In addition to computing the 3D locations, we also compute visual features for each observation o_n. These features are used to match observations over time, thereby creating 3D tracks. We denote the visual features of observation o_n as $v_n = \Psi(E_{f_n}, m_n)$, where Ψ is a function that represents the visual feature extractor applied to the mask m_n on the frame f_n.

Lifted Visual Observations. We integrate 3D locations and visual features to obtain the set of partial observations $\mathcal{W} = \{w_n : n = 1, \ldots, N\}$ in the world coordinate frame, where each w_n is a tuple (f_n, l_n, v_n). Next, we explain how these observations are matched over time to form 3D tracks.

6.3.2 Match and Keep: Matching Lifted Observations and Keeping them in Mind

In this section, we describe how we used the the set of lifted observations for associating observations with consistent identities, i.e., tracking objects over time.

We process the egocentric video E using an online approach. While using an offline approach is also an option, we opt to mimic human spatial cognition—meaning, a person becomes aware of an object's location when it is first discovered, and from that moment on the object is remembered.

Track definition. We define a track $\mathcal{T}^j$ as the collection of observations associated with a single object. The set of all tracks at time t is denoted as $\mathcal{T}_t$.

A track has one 3D location at each point in time, whether the object is in-sight or not, and we refer to the location of $\mathcal{T}^j$ at time t by $L(\mathcal{T}_t^j)$. In fact, the concept of object permanence implies that people use their spatial cognition to remember where objects are, rather than objects "disappearing" when they are obscured or move out of the egocentric camera's field of view.

The track also features a changing visual representation over time. The latter is computed at time t based on the track's most recent γ visual features' visual appearance. Limiting the average to γ recent frames takes into account the fact that objects change appearance over time—for example, a bowl may appear full, dirty, and then clean—and that older representations are less likely to support the match to future observations. The appearance of the track at time t is denoted $V(\mathcal{T}_t^j)$.

Track initialization. We initialize a new object track with an observation w_n if it represents a new, unseen object, that is not matched to another track using the online matching described next.

We define an initialization function $\mathcal{I}$, which defines the current 3D location and appearance of the observation w_n to initialize a new $\mathcal{T}^{J+1}$, where J tracks already exist. Since this is the first observation of the object, the track is projected back in time until the beginning of the video. $\forall t \leq f_n$:

$$\mathcal{I}(w_n) \rightarrow \mathcal{T}^{J+1} : L(\mathcal{T}_t^{J+1}) = l_n \text{ and } V(\mathcal{T}_t^{J+1}) = v_n \tag{6.2}$$

This reflects the common sense that objects do not magically appear out of thin air, and that an object's initial encounter indicates that it was previously there.

Track update. After initialization, a track's location and visual appearance are updated at each frame, integrating, if available, information from new observations. We define the track update function $\mathcal{U}$. It takes as input a track $\mathcal{T}^j$, the observation which will be used to update the track, and a time t. If the track $\mathcal{T}^j$ is not assigned a new observation at time t then its representation remains unchanged:

$$\mathcal{U}(\mathcal{T}^j, \varnothing, t) \rightarrow \mathcal{T}_t^j = \mathcal{T}_{(t-1)}^j \tag{6.3}$$

However, if a new observation is assigned to the track at time t, then both its location and visual feature are updated:

$$\mathcal{U}(\mathcal{T}^j, w_n, t) \rightarrow L(\mathcal{T}_t^j) = l_n \text{ and } V(\mathcal{T}_t^j) = \mu(v_n, \mathcal{T}^j) \tag{6.4}$$

where μ calculates the mean of the past γ observations assigned to the track $\mathcal{T}^j$.

Online Matching. After having obtained a set of partial 3D observations across the whole video and having defined track initialization and update functions, we now describe the online process of forming tracks from these observations. We find the set of new observations at each t; $\mathcal{W}_t = \{w_n \quad \forall n : f_n = t\}$. Note that $\mathcal{W}_t$ is empty if there are no observations at time t.

We initialize one track for each of these observations starting with the first frame in the video that has at least one observation:

$$\mathcal{T}_t = \{\mathcal{I}(w_n) \quad \forall w_n \in \mathcal{W}_t\} \tag{6.5}$$

Then, we compare $\mathcal{W}_t$ to the set of trajectories at time $t-1$ by iterating over each subsequent time. A cost function that combines visual similarity and 3D distance is used to determine matching. We follow (Rajasegaran et al., 2022) and model 3D similarity σ_L between an observation w_n and a track $\mathcal{T}^j$ by an exponential distribution, and visual similarity σ_V by a Cauchy distribution:

$$\sigma_L(w_n, \mathcal{T}^j) = \frac{1}{\beta_L} \exp\left[-D(L(\mathcal{T}_{t-1}^j), l_n)\right] \tag{6.6}$$

$$\sigma_V(w_n, \mathcal{T}^j) = \frac{1}{1 + \beta_V D(V(\mathcal{T}_{t-1}^j), v_n)^2} \tag{6.7}$$

where D is the Euclidean distance and β_L and β_V are relative weights for location and visual similarities.

We define the cost Φ of assigning an observation to an existing track as a combination of 3D and visual distance:

$$\Phi(w_n, \mathcal{T}^j) = -\log\left(\sigma_L(w_n, \mathcal{T}^j)\right) - \log\left(\sigma_V(w_n, \mathcal{T}^j)\right) \tag{6.8}$$

We use the Hungarian algorithm ξ, which computes Φ between every observation in $\mathcal{W}_t$ and the tracks $\mathcal{T}_{(t-1)}$. It returns a set of track assignments for time t, A_t, where $A_t^j = w_n$ denotes that the observation $w_n \in \mathcal{W}_t$ is to be assigned to track $\mathcal{T}^j$. A threshold for assignment cost is set to α.

$$A_t = \xi(\Phi, \mathcal{W}_t, \mathcal{T}_{t-1}) \tag{6.9}$$

We now update the set of all tracks and initialise new tracks for unassigned observations.

$$\mathcal{T}^t \leftarrow \begin{cases} \mathcal{U}(\mathcal{T}^j, A_t^j, t) & \forall j \\ \mathcal{I}(w_n) & \forall w_n \in \mathcal{W}_t : \left(\nexists j : A_t^j = w_n\right) \end{cases} \tag{6.10}$$

By using the proposed online matching, we are able to estimate the 3D location of each object for which at least one observation is available.

6.3.3 LMK for object visibility and positioning

The *Lift-Match-and-Keep* process described above facilitates spatial cognition, offering detailed insights into the visibility of each object in relation to the camera wearer at time t.

An object j can be *one* of:

- **In-sight**: if the corresponding track is assigned an observation at time t, i.e., $A_t^j \neq \varnothing$
- **Occluded**: if $L(\mathcal{T}_t^j)$ is within the field of view of the estimated camera C_t, but there is no corresponding observation ($A_t^j = \varnothing$). This might occur when an object is inside a container, such as a fridge, drawer, or cupboard, or it is occluded by the wearer's hands.
- **Out-of-view**: if $L(\mathcal{T}_t^j)$ is outside the field of view of the estimated camera C_t.

An object may also be referred to as **Out-of-sight** if it is either out-of-view or occluded (*i.e.* in the camera's viewing direction but cannot be detected).

LMK also discloses the relative distance between the object and the camera-wearer or the static environment:

- **In-reach**: if the distance from object j to the camera's position at time t is less than the approximation of the camera wearer's near space η: $D(L(\mathcal{T}_t^j), C_t) \leq \eta$
- **Out-of-reach**: as in-reach, but if $D(L(\mathcal{T}_t^j), C_t) > \eta$.
- **Moved**: object j has moved *relative to the environment* between times t_1 and t_2 if $D(L(\mathcal{T}_{t_2}^j), L(\mathcal{T}_{t_1}^j)) \geq \epsilon$, where ϵ is a minimum threshold (to account for small errors in camera and object positions).
- **Stationary**: as moved, but $< \epsilon$.

Note that the object j at time t may be both *e.g.* occluded but in-reach.

6.4 Experiments

In this section, we validate the effectiveness of the proposed method in addressing the OSNOM task. Section 6.4.1, introduces our benchmark for the OSNOM task. This evaluates the ability to determine object locations at any time given an egocentric video. Section 6.4.2 details baseline methods used for comparison. Section 6.4.3 presents the main results and qualitative examples. Section 6.4.4 ablates LMK, including its capabilities for spatial cognition.

6.4.1 Benchmarking OSNOM

Dataset. We evaluate LMK using the EPIC-KITCHENS (Damen et al., 2022) dataset. The latter contains unscripted recordings of individual participants, where all object movements in these videos result from the camera wearer interacting with and moving objects.

We extract 3D point clouds and dense camera poses using the pipeline proposed from EPIC-Fields (Tschernezki et al., 2024). For defining observations, use masks from VISOR (Darkhalil et al., 2022). For fair comparison, we use the same input for our method and baselines. In total, we evaluate on 100 videos. Those are 12 minutes long on average, and contain a total of 7.9M masks, which correspond to 2939 objects. We use the object semantic label only for calculating the ground truth for evaluation.

For most of our results, we use masks provided by VISOR, which are interpolations based on ground-truth masks. This approach enables us to assess LMK's performance without introducing errors from a detector. For completeness, we also ablate these results using a semantic-free detector (Shan et al., 2020) in Section 6.4.4. We use an additional set of 10 videos for hyperparmeter tuning.

Benchmark task. We identify a set of frames $\mathcal{F}$ where 3 or more objects are being interacted with. Each frame $f \in \mathcal{F}$ includes objects that are visible, and our goal is to evaluate the methods' ability to accurately determine the 3D locations of these objects across frames $f \pm \delta$. We assess the performance of various methods as δ increases. In total, our evaluation starts from $\mathcal{F} = 3467$ frames, with locations at 1M frames and 2171 objects, averaging 15,000 frames and 20 objects per video. Our benchmark will be made publicly available for comparisons.

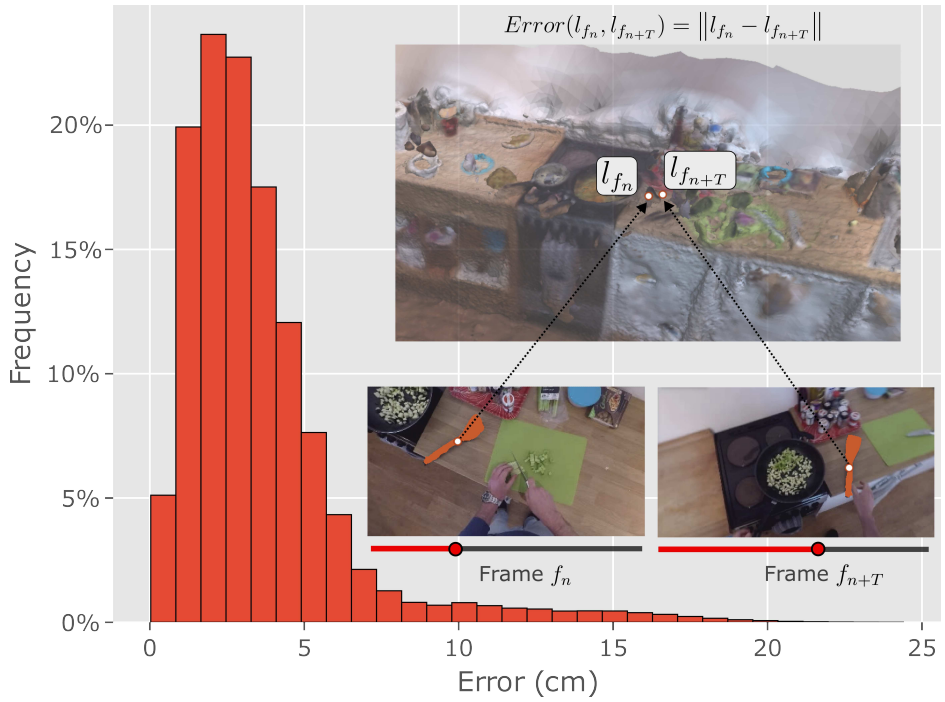


Fig. 6.4 **3D Projection error.** Distribution of projections errors in terms of Euclidean distance for the same object, at the same location, between measurements l_n to l_{n+T}.

Ground truth locations. Note that there is currently no egocentric dataset with 3D object annotations for *dynamic objects over time*. We utilize our 2D to 3D lifting approach, presented in Section 6.3.1, to establish ground-truth locations. We evaluate its accuracy in the following manner.

A random selection of objects and their corresponding time segments, during which they remain in the same location across the environment, are chosen. Although the ground-truth is unknown, comparing the errors between projections from multiple views of the same object at the same location offers an alternative method for assessing the accuracy of our 3D locations. Given multiple instances of the same object at the same location, we calculate the mean 3D error of our 3D locations. Our analysis (details provided in the supplementary material) reveals that the mean 3D error is 3.5cm, with 88% of all errors being less than 6cm and 96% of all errors less than 10cm (Figure 6.4). Based on these findings, we deem our lifting approach sufficiently precise to serve as ground-truth locations. This also shapes our metric, ensuring that our threshold for accepting assignments is significantly larger than the observed error.

Evaluation metric. We introduce a metric known as the Percentage of Correct Locations (PCL), inspired by the Percentage of Correct Keypoints (PCK) (Yang and Ramanan, 2012)

used in pose estimation evaluations, to assess the spatial alignment of objects. Traditional tracking metrics fail to evaluate tracks when objects are out of sight (Bernardin and Stiefelhagen, 2008; Luiten et al., 2021; Ristani et al., 2016). In contrast, PCL considers a predicted 3D location to be correct if its Euclidean distance from the ground truth 3D location is less than a threshold R.

For our principal experiments, we set $R = 30\,\text{cm}$[1], reflecting the idea that spatial cognition's function includes knowing an object's location with enough precision to navigate towards it or retrieve it (Downs and Stea, 1973; Waller and Nadel, 2013). R is both visualized in our experiments and ablated.

6.4.2 Experimental setup

Baselines. Since there are no prior works that have addressed the OSNOM task, we compare LMK against four baselines:

- **Random Matching**: each observation is randomly assigned either to an existing track or to initiate a new track, illustrating the complexity of the data.
- **Out of Sight, Lost (OSL)**: objects are considered "lost" when they move out-of-view, at which point the PCL is reported as 0, and their tracks are terminated. This baseline emphasizes the inherent challenge in egocentric video analysis, where objects frequently move out of the camera's view shortly after being observed.
- **Out of Sight, Out of Mind (OSOM)**: observations can only be assigned to tracks that are currently in view. Once a track goes out-of-view, the PCL is reported as 0, and the tracks are frozen until they re-enter the camera's field of view. This scenario represents an upper bound for tracking accuracy within the camera coordinate frame.
- **EgoLoc** (Mai et al., 2022): we adapt this state-of-the-art VQ3D approach to the OSNOM task to manage multiple objects. We utilize the same masks, features, 3D scene, and lifting technique for a fair and direct comparison. EgoLoc's method of weighted averaging over all past observations is not suitable for OSNOM due to objects changing positions; instead, we opt for the most *recent* match.

Implementation details. For appearance features, Ψ, we utilize a DINO-v2 (Oquab et al., 2023) pre-trained model. We crop each mask, scale it to 224×224, and then pass it to the

[1] This is half the standard width of a cupboard or cabinet, which is 60 cm or 24 inches.

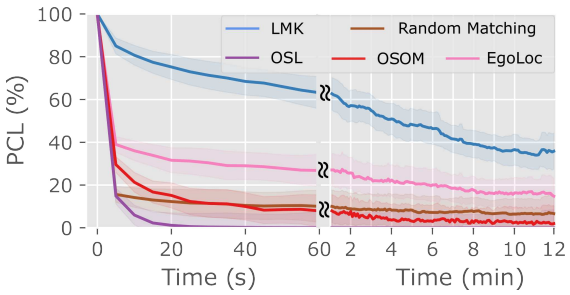


Fig. 6.5 **OSNOM results.** PCL results of LMK compared to baselines. Results are shown from 0-60 seconds, then 1-12 minutes.

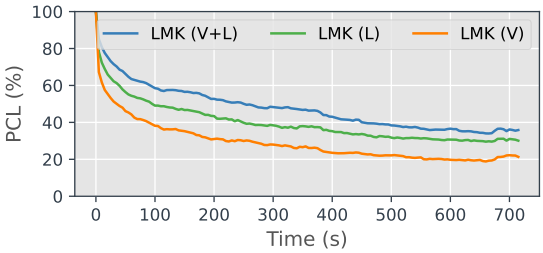


Fig. 6.6 **Effect of visual appearance and location**. PCL results of LMK for visual features (V), location features (L), or both (V+L).

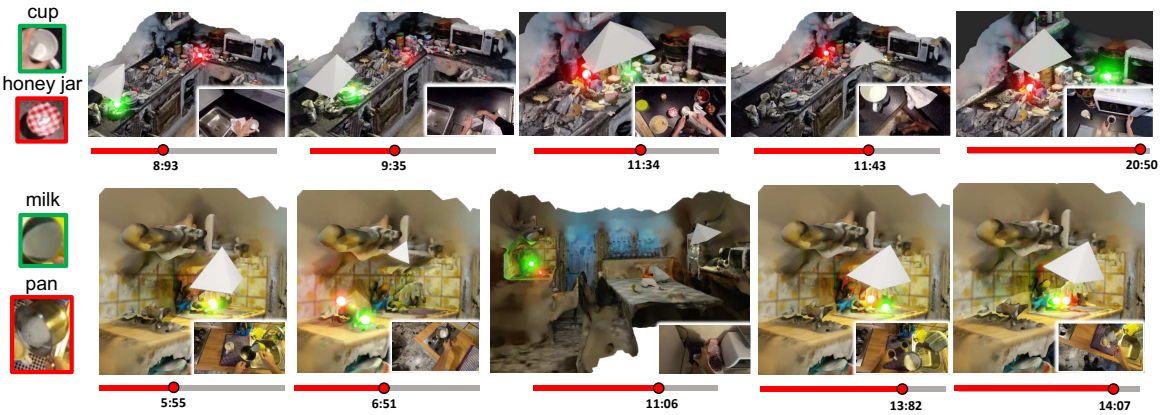


Fig. 6.7 **3D location prediction.** Predicted 3D locations (neon dots) of two objects (left) across multiple frames, with insets showing each frame (right). Note how the object locations are accurately maintained, even when the camera-wearer is at a distance (bottom middle).

backbone. We ablate the choice of features. We set $\alpha = 10$, $\gamma = 100$, $\beta_L = 13$, and $\beta_V = 2$ (chosen based on validation set performance). Meshes are computed in advance, requiring an average of 5 hours per video on one 2080Ti. For online tracking, DINOv2 operates at 30 FPS and lifting-to-3D processes at 200 FPS on one P100. LMK runs at 1000 FPS on a single CPU core.

6.4.3 Results

Results for the OSNOM task, comparing LMK against the baselines of OSL, OSOM, Random Matching, and EgoLoc, are shown in Figure 6.5. We report the average PCL (on the y-axis) across the entire dataset for each 5-second evaluation interval (x-axis), with the standard deviation shown as a shaded area. Performance is evaluated over both short-term (0-60 seconds) and long-term (1-12 minutes) intervals. As time progresses, the complexity of matching observations increases due to an increase in the number of objects being interacted with and tracked. Consequently, a drop in performance over time is observed for all methods.

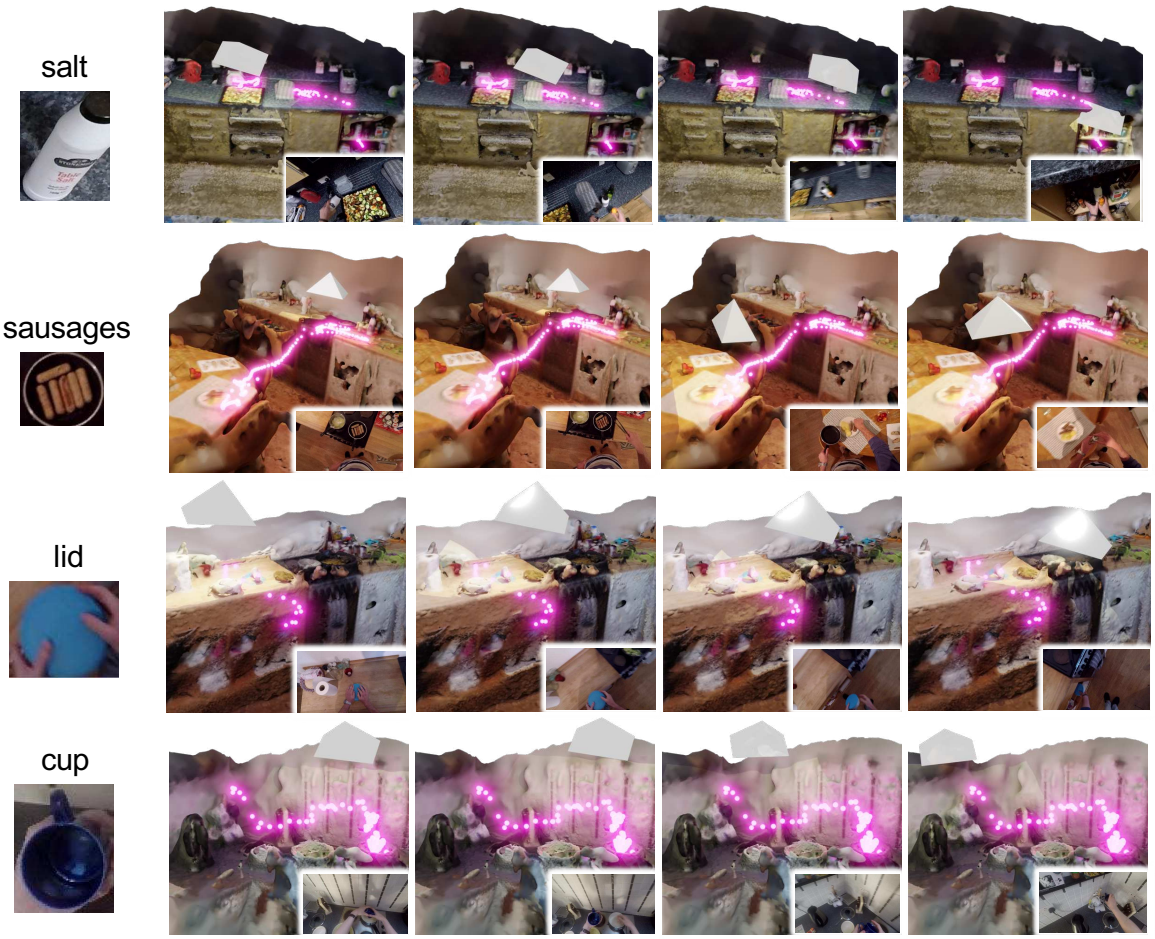


Fig. 6.8 **Trajectory prediction** for objects in motion. Neon dots represent the predicted 3D positions along with corresponding camera poses. Objects are accurately positioned, whether they are stationary (resting on surfaces) or moving (carried in-hand).

LMK demonstrates significant improvements over all baselines. In comparison to EgoLoc, it shows a 39% average improvement in tracking up to 1 minute, and 25% improvement from 1 to 12 minutes. This performance boost is attributed to LMK's tracking across consecutive frames, enhancing its robustness against variations in object appearance caused by changes in orientation or occlusion, and its utilization of 3D locations for matching.

The pronounced decline in performance observed with the OSOM and OSL baselines highlights the challenges presented by egocentric footage, where the constant movement of the camera wearer frequently causes objects to exit the field of view. Specifically, when tracking is limited to objects while they are in view (OSL baseline), performance drops to zero shortly after 20 seconds, indicating rapid loss of objects from sight. The OSOM baseline, which considers only objects within the camera's field of view without accounting for 3D world coordinates and object permanence, proves to be inadequate for the OSNOM task, performing even worse than random assignment.

Qualitative results. Figure 6.7 shows the predicted locations of several objects at discrete time intervals. In Figure 6.8, we present the 3D trajectories of objects as they are manipulated by the camera wearer. For instance, we depict the trajectory of the *salt bottle* from being in hand (pouring salt), to being placed on the countertop, and eventually being returned to a lower cupboard, while the *cup* ends up on a hanger. In all instances, LMK successfully tracks objects both when they are static (on surfaces) and when they are in motion (in-hand).

6.4.4 LMK Ablation

Effect of visual appearance and location. LMK assigns observations to tracks based on appearance and location similarities. Figure 6.6 illustrates the impact of relying solely on visual appearance (V) and solely on location (L) compared to the default combination of both (V+L). This combination yields improvements (mean +19% over V, +8% over L), underscoring that appearance and location are complementary attributes. Appearance is good in frame-to-frame assignments, while location proves particularly useful for tracking objects in motion, those occluded, and for reassigning objects when they re-enter the field of view.

Accuracy at different radii. In all our experiments, we set the PCL threshold to $R = 30\,\text{cm}$. Figure 6.9 also presents results for when this threshold is increased to $R = 60\,\text{cm}$ and $R = 90\,\text{cm}$, which are visualized in 3D to illustrate their respective challenges. As expected, the PCL value increases with larger R values.

Visual features. Our default feature extractor Φ is a ViT (Dosovitskiy et al., 2020), pre-trained using the self-supervised DINO-v2 approach (Oquab et al., 2023). We also explore ViTs pre-trained on CLIP (Radford et al., 2021) and ImageNet (Deng et al., 2009), as shown in Figure 6.10. DINO-v2 surpasses the other methods across all time scales, likely because the pre-training tasks of CLIP (vision and language alignment) and ImageNet (image classification) are less aligned with our need for consistent frame-to-frame visual similarity.

Detections. We utilized annotations from VISOR (Darkhalil et al., 2022) for 2D masks to avoid compounding detection errors when assessing the accuracy of 3D location estimation, which is our primary focus. In Figure 6.11, we present an ablation study using detections from (Shan et al., 2020). This model generates semantic-free bounding boxes for actively interacting objects, which are then used as inputs for LMK and the best-performing baseline, EgoLoc. For result evaluation, we match each detection with the VISOR object that achieves the highest Maximum Intersection Over Union (MIOU). LMK continues to significantly outperform EgoLoc.

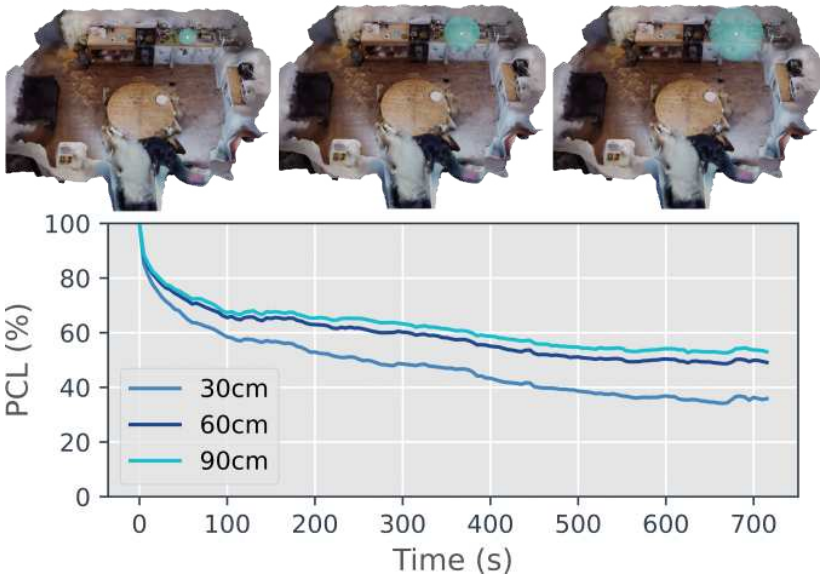


Fig. 6.9 **Evaluation thresholds.** LMK results when increasing the PCL threshold R, which is the maximum distance between predicted and ground truth 3D locations deemed successful. Visualizations display the regions encompassed by volumes of $R = 30\,\text{cm}, 60\,\text{cm}$, and $90\,\text{cm}$ in blue, centered on the counter.

Weighting visual appearance and location. LMK employs the hyperparameters β_V (see Eq. 6.6) and β_L (see Eq. 6.7) to adjust the importance of visual and location similarities when assigning new observations to tracks. These hyperparameters are selected based on optimal performance on the validation set, averaged across different timescales. Figure 6.12a illustrates the validation set performance when $\beta_V = 2$ is fixed and β_L is varied. Conversely, Figure 6.12b demonstrates the performance when $\beta_L = 13$ is fixed and β_V varies. Both hyperparameters exhibit relative stability, likely attributed to their scaling by appropriate distributions (Cauchy and Exponential).

Track visual appearance history. Figure 6.12c ablates γ over the validation set, where γ represents the number of recent features averaged for the visual representation. Optimal results are achieved with $\gamma = 100$, showing diminished performance for both smaller and larger values of γ, yet the performance remains relatively stable down to a single observation. A low γ value, indicating insufficient accumulation of appearance information, fails to ensure a consistent representation of objects over time. This is particularly problematic in egocentric videos, where varying perspectives and partial occlusions are common. Conversely, aggregating appearance features over long periods (high γ) can lead to feature inconsistencies, especially for objects whose appearance changes a lot over time.

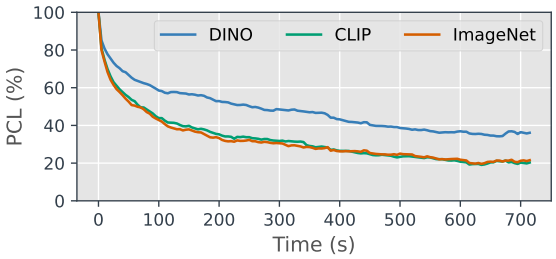


Fig. 6.10 **Visual feature choice** of a DINO-v2, CLIP or ImageNet (ViT).

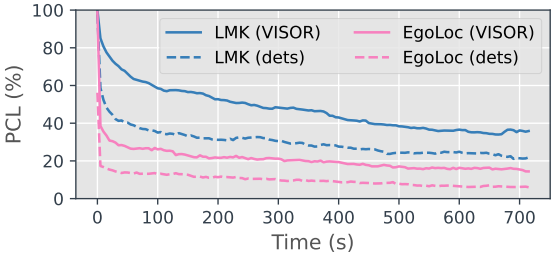
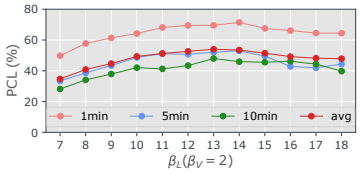
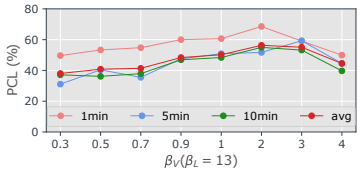


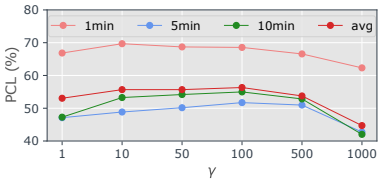
Fig. 6.11 **Detections.** LMK on both visual and location features when using VISOR annotations *vs* using detections from (Shan et al., 2020).



(a) β_L, the weighting of 3D location for assigning observations to tracks.



(b) β_V, the weighting of visual appearance for assigning observations to tracks.



(c) γ, the number of visual appearance features averaged to calculate track representation.

Fig. 6.12 **Hyperparameter ablations** for LMK on the validation set. We choose the best average over 1, 5 and 10 minute sequence lengths.

LMK for spatial cognition. Figure 6.13 illustrates the performance of LMK on object states as defined in Section 6.3.3. For each state combination of (In-reach[2], Out-of-reach) and (In-sight, Occluded, Out-of-view), we report the total number of ground truth objects and the number LMK accurately locates over a 1-minute interval. Despite objects being interacted with for over 1 minute, LMK continues to accurately determine their locations, achieving an average accuracy of 72%. Moreover, LMK achieves an accuracy of 82% for objects that are both out-of-reach and out-of-view.

Effect of objects going out-of-view. We analyze the effect of a track disappearing from view and then reemerging within a 10-minute span, as depicted in Figure 6.14. The LMK method, which leverages 3D locations for matching, shows a substantial improvement in performance under these circumstances.

Moved *vs.* Stationary objects. Figure 6.15 presents the performance of the PCL (Point Cloud Localization) method when applying a movement threshold of $\epsilon = 30$cm. The results indicate that the tracking accuracy for stationary objects is, on average, 35% higher than for objects that have been moved. This discrepancy can be attributed to the fact that objects

[2]A reachable threshold of $\eta = 70$cm is used.

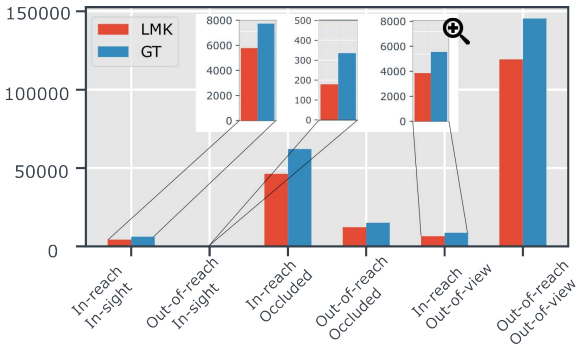


Fig. 6.13 **LMK for spatial cognition.** Number of objects correctly located by LMK, separately by combinations of (In-reach, Out-of-reach) and (In-sight, Occluded, Out-of-view).

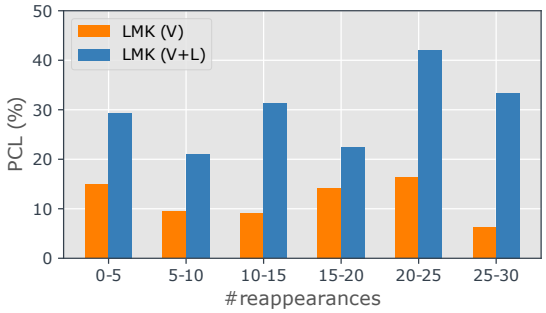


Fig. 6.14 **Effect of reappearing.** Evaluation is performed over 10 minutes, for LMK with visual appearance (V) and the combination of visual appearance and location (V+L).

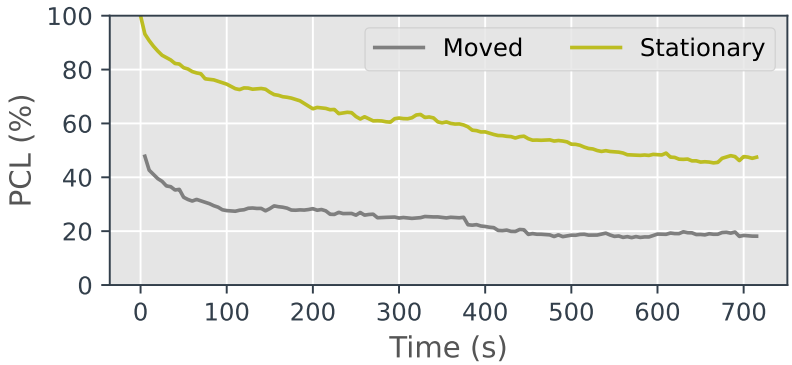


Fig. 6.15 LMK Results for **Moved vs Stationary** objects with respect to the environment.

often appear visually different after being moved, for example, due to changes in orientation or lighting conditions.

Failure cases. We identify two primary reasons for failure cases in the LMK method. For clarity, we illustrate each case separately in Figure 6.16 and Figure 6.17. In each figure, we focus on a single object and depict its predicted trajectory in green. Predictions that fail are indicated in red, where we plot the accurate ground truth trajectory.

In Figure 6.16, we present cases where the tracking is momentarily lost but subsequently correctly reacquired. In the first scenario, a tin is accurately tracked for the majority of its journey, including when it is discarded in the bin. However, for a brief period, the predictions are incorrect, as highlighted by the red dots. Similarly, in the second scenario, a knife is inaccurately predicted while it is obscured by a hand or when in hand. The final example highlights failures in predicting the correct trajectory of a pot as it is filled with milk, which alters its appearance. Coincidentally, the pot is moved out of the camera's

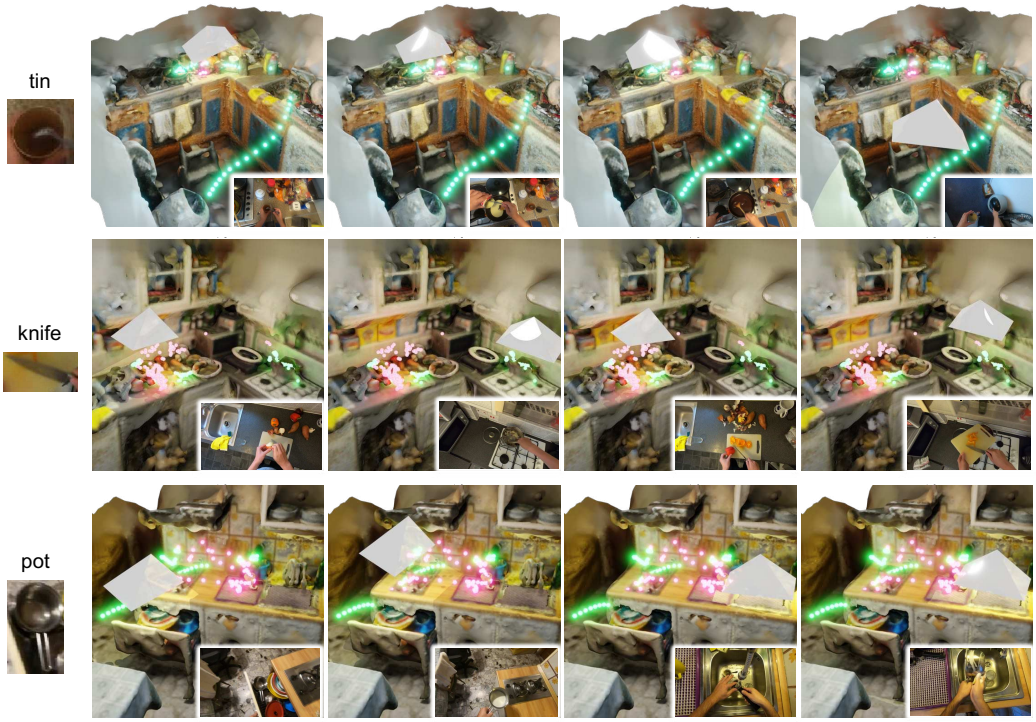


Fig. 6.16 **Trajectory prediction - temporarily lost but recovered track.** Predicted trajectory of three objects in motion. Green neon dots represent accurately predicted 3D positions across four frames along with their corresponding camera views, while red neon dots indicate the ground-truth trajectory where predictions fail. Although tracking momentarily fails, the object is accurately matched to a future observation shortly afterward.

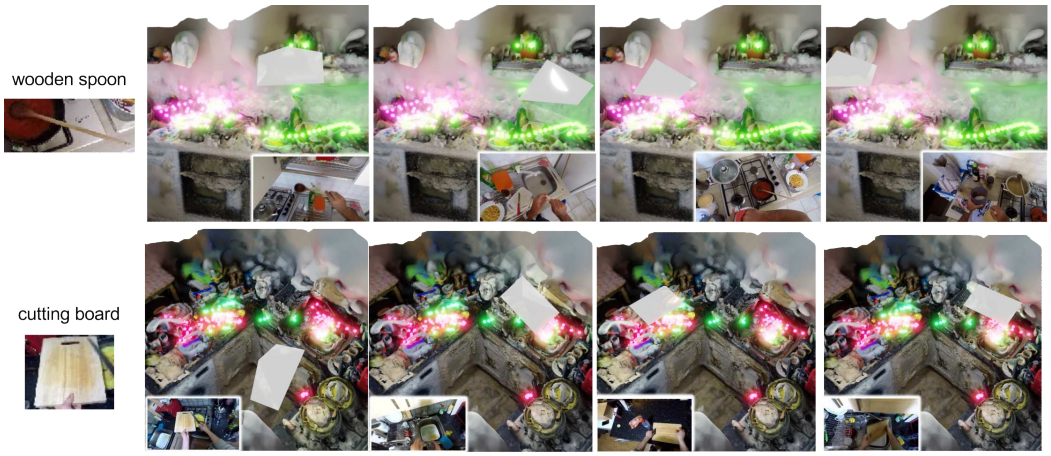


Fig. 6.17 **Trajectory prediction - lost track.** Predicted trajectory of two objects in motion. Green neon dots indicate correctly predicted 3D positions across four frames along with their corresponding camera views, while red neon dots display the ground-truth trajectory where predictions fail. When tracking fails, all subsequent predictions are assigned to a new track.

view, resulting in failures in both appearance and location matching. However, as the pot is emptied, appearance matching improves towards the end of the track.

In Figure 6.17, we demonstrate failure cases where tracking is not recovered. In the first example, a wooden spoon is attributed a new trajectory, and tracking continues under a new identity. A similar situation arises for a cutting board when it is moved to a cluttered sink. Failures predominantly occur in cluttered settings, such as when slicing peppers with a knife in Figure 6.16, or when stirring with a spoon in Figure 6.17. In these instances, the proximity of multiple objects leads to overlapping locations, making the individual object's location less useful for matching.

6.5 Conclusion

In this chapter, we introduce the task of "Out of Sight, Not Out of Mind" (OSNOM) for egocentric videos with partial object observations. This task evaluates the 3D tracking performance of active objects, both when they are visible and when they are out of sight. To address this task, we merged partial 2D-based information with comprehensive 3D information about the location of objects in the scene where the videos are recorded. We presented Lift, Match, and Keep (LMK), a method that lifts partial 2D observations from camera coordinates to 3D world coordinates, matches them over time using visual appearance and 3D location, and keeps track of them even when they disappear from view. Results from long-duration videos in the EPIC-Kitchens dataset show that LMK achieves promising results for both short-term (64% accuracy up to 1 minute) and long-term (37% accuracy for 1-12 minutes) periods. These findings highlight the significance of maintaining 3D world locations for objects that go out of view. For future work, we aim to explore whether LMK can effectively track objects through state changes and investigate the potential for shared 3D object tracks between multiple ego- and exo-centric cameras. A future direction involves expanding the OSNOM task to multiple videos, over time. This aligns with our ultimate goal of developing an assistive solution that maintains awareness of object locations over hours and potentially days.

Chapter 7

Conclusions and future works

In this thesis, we investigated how egocentric video representations might benefit from multi-modal data. In the first part of the thesis, we demonstrated how solving auxiliary tasks across multiple information channels can improve the models' generalization capabilities in cross-domain scenarios. We then introduced new types of data within the context of egocentric vision, namely event-based data and 3D information. We analyzed the challenges associated with integrating these data types with standard RGB information, along with their respective advantages and disadvantages. In this chapter, we summarize the main takeaways, limitations, and future works of each chapter.

Multi-Modal Relative Norm Alignment for Tackling the Domain Shift

In Chapter 3, we introduced a method aimed at addressing the challenge of multi-modal Domain Generalization (DG). Our methodology draws from the observation that differences in the marginal distributions of modalities can significantly affect the training process, leading to variances in their feature norms. This ultimately leads to sub-optimal performance in cross-domain scenarios. Starting from this observation, we introduced the Relative Norm Alignment (RNA) loss, which is designed to equalize the feature norms extracted from various modalities. We show that re-balancing the contribution of different modalities during training improves the overall model accuracy in cross-domain scenarios. We demonstrated how this loss can be seamlessly integrated into Unsupervised Domain Adaptation (UDA) scenarios, where it works in conjunction with an adversarial loss and Information Maximization to enhance feature transferability on the target domain.

Our experiments involved the integration of visual, audio, and optical flow data. Future research could explore applying this approach to different modalities, where the issue of heterogeneity may be more pronounced. For instance, merging visual and textual information presents an interesting avenue for exploration. Although our current focus was on the activity recognition task, future efforts could extend to other tasks and to different model architectures, broadening the applicability of our approach. Finally, RNA rebalances the contribution of different modalities at feature level. A potential variation to explore might be modulating the backpropagated gradient, similarly to what has been recently proposed in (Wu et al., 2022b). This adjustment could provide a different perspective on addressing the modality imbalance by directly influencing the learning process.

A limitation we observed arises from the fact that in many real-world cases, data distributions are strongly unbalanced, leading to lower accuracy for the tail classes (Buda et al., 2018). The literature shows how this imbalance results in unbalanced norms of classification weights per class (Guo and Zhang, 2017; Kim and Kim, 2020), as well as unbalanced norms of features per class (Li et al., 2022a; Wu et al., 2017). In developing our method, we hypothesized that balancing the norms per class could positively affect the rebalancing of the classifier's weights for the tail classes. However, our experimental results did not demonstrate this effect. This opens up possibilities for future developments to incorporate this objective into RNA as an additional component that rebalances the weights of the classifier.

Vision and Language for Domain Generalization

In Chapter 4 we use textual information to address the issue of generalization across varying scenarios and locations, positing that it is possible to learn actions in a way that enables their recognition across new contexts (e.g., identifying the action "cut" in cooking as analogous to "cut" performed by a mechanic) and geographical settings (e.g., recognizing the action "cut" in Italy as the same action in India). This concept forms the core motivation of our research.

To address this challenge, we introduced ARGO1M, a curated dataset tailored for this purpose. In response to the complex task of adapting to diverse scenarios and locations, we developed a novel method grounded in a visual-text reconstruction task. This technique involves reconstructing videos from a combination of videos from different scenarios and locations using text narrations to guide reconstructions. This approach ensures that the reconstructions are not biased toward visual domain-specific information but are instead informed by semantically related data. Our approach demonstrates superior performance over existing baselines, as validated by extensive analysis and in-depth ablation studies.

The complexities posed by ARGO1M represent a significant advancement in domain generalization research. We hope that this chapter will inspire further research in domain generalization, particularly in video analysis, an area ripe for deeper exploration. Future research directions could focus on generalizing to more complex actions, particularly those involving combinations of verbs and nouns. Considering different hierarchies at the fine-grained level for classifying actions could be advantageous. For instance, in our study, actions like "trimming" were grouped under the broader category of "cutting". However, alternative grouping strategies could be explored. Additionally, exploring the zero-shot learning capabilities of Large Language Models (LLMs) presents an opportunity to extend the model's applicability beyond the predefined set of 60 actions. By leveraging these capabilities, we can envision a framework where the model recognizes actions in a zero-shot setting, transcending the limitations of the initially proposed action set. Finally, since the reconstruction is currently performed at the feature level, it might be interesting to consider reconstructing at pixel level. This shift could provide a more granular understanding of the visual components of the actions, allowing for a deeper analysis of how different elements interact and contribute to the overall action recognition.

Event-based Data for Egocentric Vision

In Chapter 5 we introduced N-EPIC-Kitchens, a pioneering dataset for event-based egocentric action recognition. We conducted a comprehensive comparative analysis to evaluate the performance of event-based data against traditional RGB and optical flow information. This study aimed to highlight the strengths and limitations of each data type across various application scenarios. We proposed and evaluated two novel methodologies specifically designed for event data–E^2(GO) and E^2(GO)MO–tailored at exploiting the temporality encoded by event-based information. These methods leverage the unique characteristics of event data to enhance the understanding and processing of dynamic scenes.

With the introduction of event data into the domain of egocentric action recognition, our objective is to facilitate a direct comparison with established benchmarks in the field, such as those proposed by Damen et al. (2018), and further scaled by Damen et al. (2022), positioning the event modality as a competitor against traditional modalities. This goal led us to favor simulation of event data over the creation of a new first-person dataset. Our decision is validated by an in-depth analysis of the Sim-to-Real gap, where we demonstrate that, through the application of traditional Unsupervised Domain Adaptation techniques, simulated event data can effectively generalize to real-world scenarios. Future research could involve recording a new dataset that includes both RGB and event data, utilizing recent

advancements in event-based cameras that provide both modalities (Berner et al., 2013). This would allow us to test the differences in our models, which were pre-trained on simulated data, when applied to both simulated and real data. Such a study would offer valuable insights into the adaptability and performance of our models in various real-world scenarios.

Moreover, our findings reveal that, despite the significant computational and temporal overheads, the TV-L1 optical flow algorithm exhibits exceptional performance, particularly in terms of its resilience to domain shifts. We ascribe this superiority primarily to the algorithm's capability to filter out camera motion, thereby providing a more refined motion analysis compared to that offered by raw event data. In fact, we observed that the motion of the camera itself inadvertently captures event information near objects in the background. Future research directions could include the exploration of motion compensation strategies, which are frequently utilized in event data processing (Stoffregen et al., 2019), to further mitigate background noise and enhance data utility.

Finally, directly extracting optical flow from event data, rather than distilling it from the optical flow computed from the RGB stream, could be a promising approach. This method might enhance the efficiency and accuracy of capturing motion dynamics by leveraging the high temporal resolution of event data. Such an approach could significantly improve the detection and analysis of movement within a scene.

Egocentric Video Understanding using 3D

In Chapter 6, we explored the integration of 3D information about the scene with 2D frame-based information extracted from traditional RGB egocentric cameras. To underscore the importance of 3D information, we introduced the task of "Out of Sight, Not Out of Mind" (OSNOM) in the context of egocentric videos that feature partial object observations. This task emphasizes the significance of 3D data in enhancing our understanding and interpretation of scenes where objects are only partially visible, demonstrating how 3D information can complement 2D imagery to provide a more complete and contextually rich analysis. We evaluated the capabilities of tracking active objects in 3D, focusing on their behavior both when they are within the visual field and when they temporarily disappear from view. To address this challenge, we proposed a novel approach named *Lift, Match, and Keep* (LMK). This method *lifts* partial 2D observations from camera coordinates into 3D world coordinates, *matches* these observations across time by leveraging visual appearance and spatial location, and *keeps* a continuous track of them even when they are not visible. Empirical results on long-duration videos from the EPIC-Kitchens dataset indicate that LMK

achieves promising performance, achieving 64% accuracy for tracking durations up to one minute and maintaining 37% accuracy for tracking from one to twelve minutes. These results highlight the critical role of preserving 3D world locations for objects when they move out of the camera's field of view.

Looking ahead, future works could investigate LMK's capability to track objects through their state changes or when they divide into multiple parts. For example, when opening an egg, the shell and the contents become two distinct entities that could be tracked separately. This exploration would extend LMK's applicability to scenarios where objects undergo significant transformations, providing deeper insights into complex dynamic processes.

Moreover, interesting future studies could assess the feasibility of integrating shared, 3D object tracks across multiple egocentric and exocentric camera perspectives (Grauman et al., 2023). This would enable a comprehensive multi-view analysis, enhancing object tracking accuracy and robustness by synthesizing information from various angles and viewpoints. Such an approach could significantly improve spatial awareness and object interaction understanding in complex environments.

Moreover, the scope of the OSNOM task could be broadened to encompass multiple video sequences over extended periods. By utilizing initial assumptions about object locations from previous observations as priors, OSNOM's applicability can be enhanced. Extending our focus beyond single video analysis aligns with our ultimate objective of developing an assistive system that maintains awareness of object placements over hours or potentially days, thereby providing a more holistic understanding of object dynamics in everyday environments.

References

Blender. URL https://www.blender.org. Accessed: Feb. 24, 2021. 125

Nakul Agarwal, Yi Ting Chen, Behzad Dariush, and Ming Hsuan Yang. Unsupervised domain adaptation for spatio-temporal action localization. In *31st British Machine Vision Conference, BMVC 2020*, 2020. 36

Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva. With whom do i interact? detecting social interactions in egocentric photo-streams. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2959–2964. IEEE, 2016. 19

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019. 57, 59

Stefano Alletto, Giuseppe Serra, Simone Calderara, Francesco Solera, and Rita Cucchiara. From ego to nos-vision: Detecting social relationships in first-person views. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 580–585, 2014. 19

Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 51

Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7243–7252, 2017. xiv, 53, 111, 112

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 20, 24

Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018. 41

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019. 55

Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022. 26, 28

Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018. 28

Francesco Barbato, Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Latent space regularization for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2835–2845, June 2021. 65

Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022. 29

Francisco Barranco, Cornelia Fermuller, Yiannis Aloimonos, and Tobi Delbruck. A dataset for visual navigation with neuromorphic methods. *Frontiers in neuroscience*, 10:49, 2016. 51

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 86, 87

Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. *IEEE transactions on neural networks and learning systems*, 25 (2):407–417, 2013. 49

Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *International Conference on Computer Vision*, 2019. 150

Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008. 159

Raphael Berner, Christian Brandli, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 10mw 12us latency sparse-output vision sensor for mobile applications. In *2013 Symposium on VLSI Circuits*, pages C186–C187. IEEE, 2013. 46, 171

Gedas Bertasius, Hyun Soo Park, and Jianbo Shi. Exploiting egocentric object prior for 3d saliency detection. *arXiv preprint arXiv:1511.02682*, 2015. 31

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021. 24

Yin Bi, Aaron Chadha, Alhabib Abbas, Eirina Bourtsoulatze, and Yiannis Andreopoulos. Graph-based object classification for neuromorphic vision sensing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 491–501, 2019. 53

O Blender. Blender—a 3d modelling and rendering package. *Retrieved. represents the sequence of Constructs1 to*, 4, 2018. 128

János Botzheim, Takenori Obo, and Naoyuki Kubota. Human gesture recognition for robot partners by spiking neural network and classification learning. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 1954–1958. IEEE, 2012. 48

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 134

Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *International Conference on Computer Vision*, 2009. 149

John Bridle, Anthony Heading, and David MacKay. Unsupervised classifiers, mutual information and 'phantom targets. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. 67

Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio Maria Carlucci, Barbara Caputo, and Tatiana Tommasi. Self-supervised learning across domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5516–5528, 2021. 36, 40, 41

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.*, 106:249–259, oct 2018. ISSN 0893-6080. 80, 169

Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10(12):551–557, 2006. 147

Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. Asynchronous convolutional networks for object detection in neuromorphic cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 114

Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 136–152. Springer, 2020a. 48

Marco Cannici, Marco Ciccone, Andrea Romanoni, and Matteo Matteucci. A differentiable recurrent surface for asynchronous event-based data. In *European Conference on Computer Vision*, pages 136–152. Springer, 2020b. 114, 133

Marco Cannici, Chiara Plizzari, Mirco Planamente, Marco Ciccone, Andrea Bottino, Barbara Caputo, and Matteo Matteucci. N-rod: a neuromorphic dataset for synthetic-to-real domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2021. 8

Congqi Cao, Yifan Zhang, Yi Wu, Hanqing Lu, and Jian Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE international conference on computer vision*, pages 3763–3771, 2017. 20

Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019. 40, 41, 98

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 20, 23, 90, 91, 99, 113, 114, 118

Alejandro Cartas, Jordi Luque, Petia Radeva, Carlos Segura, and Mariella Dimiccoli. Seeing and hearing egocentric actions: How much can we learn? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 59

Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7354–7362, 2019. 36

Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7119–7129, 2022a. 41

Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. *Advances in Neural Information Processing Systems*, 35:33302–33315, 2022b. 41

Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021. 73

Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6321–6330, 2019. 4, 36, 37, 38, 67, 68, 86, 87

Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 36

Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020a. 38, 40, 87

Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020b. 36

Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020c. 38, 40

Gregory Kevin Cohen. *Event-Based Feature Detection, Recognition and Classification*. Theses, Université Pierre et Marie Curie - Paris VI ; University of Western Sydney, September 2016. 114

Giorgia Committeri, Gaspare Galati, Anne-Lise Paradis, Luigi Pizzamiglio, Alain Berthoz, and Denis LeBihan. Reference frames for spatial cognition: different brain areas are involved in viewer-, object-, and landmark-centered judgments about object location. *Journal of cognitive neuroscience*, 16(9):1517–1535, 2004. 147

Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019. 107, 109

Daniel Czech and Garrick Orchard. Evaluating noise filtering for event-based asynchronous change detection image sensors. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 19–24. IEEE, 2016. 48

Victor G Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1181–1190, 2022. 38, 39

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34: 3965–3977, 2021. 120

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 19

Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014. 31

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. xix, 4, 27, 68, 71, 74, 82, 84, 109, 111, 114, 124, 144, 170

Dima Damen, Will Price, Evangelos Kazakos, Antonino Furnari, and Giovanni Maria Farinella. Epic-kitchens - 2019 challenges report. https://epic-kitchens.github.io/Reports/EPIC-Kitchens-Challenges-2019-Report.pdf, 2019. xi, 58, 107, 109

Dima Damen, Evangelos Kazakos, Will Price, Jian Ma, and Hazel Doughty. Epic-kitchens-55 - 2020 challenges report, 2020. xi, 57, 58, 70, 107, 109

Dima Damen, Adriano Fragomeni, Jonathan Munro, Toby Perrett, Daniel Whettam, and Michael Wray. Epic-kitchens-100- 2021 challenges report. https://epic-kitchens.github.io/Reports/EPIC-KITCHENS-Challenges-2021-Report.pdf, 2021. 107, 109

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33–55, 2022. xix, xx, 26, 27, 33, 68, 69, 70, 79, 87, 90, 114, 124, 144, 148, 149, 157, 170

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 27, 149, 157, 162

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010. 107, 109

Pierre de Tournemire, Davide Nitti, Etienne Perot, Davide Migliore, and Amos Sironi. A large scale event-based detection dataset for automotive. *arXiv preprint arXiv:2001.08499*, 2020. 111

Tobi Delbruck. Neuromorphic vision sensing and processing. In *2016 46Th european solid-state device research conference (ESSDERC)*, pages 7–14. IEEE, 2016. 107, 109

Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6713–6719. IEEE, 2019. 51

Andong Deng, Taojiannan Yang, and Chen Chen. A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20519–20531, 2023. 28

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 53, 162

Wanxia Deng, Zhuo Su, Qiang Qiu, Lingjun Zhao, Gangyao Kuang, Matti Pietikäinen, Huaxin Xiao, and Li Liu. Deep ladder reconstruction-classification network for unsupervised domain adaptation. *Pattern Recognition Letters*, 152:398–405, 2021. 39

Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IEEE Robotics and Automation Letters*, 5(3):4596–4603, 2020a. 48

Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IEEE Robotics and Automation Letters*, 5(3):4596–4603, 2020b. 114

Zhijie Deng, Yucen Luo, and Jun Zhu. Cluster alignment with a teacher for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9944–9953, 2019. 36

Peter U Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International joint conference on neural networks (IJCNN)*, pages 1–8. ieee, 2015. 48, 51

Tien Do, Khiem Vuong, and Hyun Soo Park. Egocentric scene understanding via multimodal spatial rectifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2832–2841, 2022. 32

Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993, 2020. 87

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020. 20, 24, 162

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32:6450–6461, 2019. 40, 41

Roger M Downs and David Stea. *Image and environment: Cognitive mapping and spatial behavior*. Transaction Publishers, 1973. 147, 159

Wenbin Du, Yali Wang, and Yu Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 3725–3734, 2017. 21

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003. 22

Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012a. 19

Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 314–327. Springer, 2012b. 20, 26, 124, 144

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006. 52, 136

Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017. 22

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 20, 23, 91, 99

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 87

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 41

Qichen Fu, Xingyu Liu, and Kris Kitani. Sequential voting with relational box fields for active object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2374–2383, 2022. 28

Antonino Furnari and Giovanni Maria Farinella. 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*, pages 6252–6261, 2019. 26, 27

Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017. 26

Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009. 152

Guillermo Gallego, Jon EA Lund, Elias Mueggler, Henri Rebecq, Tobi Delbruck, and Davide Scaramuzza. Event-based, 6-dof camera tracking from photometric depth maps. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2402–2412, 2017. 48

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020a. 107, 109

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020b. 3, 134

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015a. PMLR. x, 36, 37, 67

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, 07–09 Jul 2015b. PMLR. 131, 137, 138, 142, 143

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 40, 72, 74, 98

Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019a. 114, 133

Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019b. 48, 51, 125, 136

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 51, 54, 112, 123, 125, 127, 135, 136, 139, 140, 142

Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3): 4947–4954, 2021. 111

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations*, 2018. 119

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 597–613. Springer, 2016. 39

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 41

Shane Gilroy, Edward Jones, and Martin Glavin. Overcoming occlusion in the automotive environment—a review. *IEEE Transactions on Intelligent Transportation Systems*, 22(1): 23–35, 2019. 150

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021. 26, 149

Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–980, 2017. 22

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 25, 27

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 29

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision*, 2023. 150

Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6481–6491, 2023. 78, 80

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 41

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 57, 59

Yves Grandvalet and Y. Bengio. Semi-supervised learning by entropy minimization. In *Adv. Neural Inform. Process. Syst.*, volume 367, pages 281–296, 01 2004. 132, 137, 138, 142, 143

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 26, 28, 29, 30, 31, 32, 85, 88, 99, 149, 153

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *Computer Vision and Pattern Recognition*, 2023. 29, 32, 172

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773, 2012. 40, 98

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 53

Xiao Gu, Jianing Qiu, Yao Guo, Benny Lo, and Guang-Zhong Yang. Transaction: Icl-sjtu submission to epic-kitchens action anticipation challenge 2021. *arXiv preprint arXiv:2107.13259*, 2021. 27

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020. 98

Yandong Guo and Lei Zhang. One-shot face recognition by promoting underrepresented classes. *arXiv preprint arXiv:1707.05574*, 2017. 80, 169

Yunzhe Hao, Xuhui Huang, Meng Dong, and Bo Xu. A biologically plausible supervised learning method for spiking neural networks using the symmetric stdp rule. *Neural Networks*, 121:387–395, 2020. 48

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017. 20, 23

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 23, 118, 136

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022. 120

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 36

Tsubasa Hirakawa, Takayoshi Yamashita, Toru Tamaki, and Hironobu Fujiyoshi. Survey on vision-based path prediction. In *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts: 6th International Conference, DAPI 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part II 6*, pages 48–64. Springer, 2018. 2

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997. 133

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. 90

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021. 41

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 115

Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 10: 405, 2016. xiv, 51, 53, 111, 112, 135

Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22910–22921, June 2023. 29

Yan Huang and Irfan Essa. Tracking multiple objects through occlusions. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 1051–1058. IEEE, 2005. 149

Thomas Huckle and Alexander Kallischko. Frobenius norm minimization and probing for preconditioning. *International Journal of Computer Mathematics*, 84(8):1225–1248, 2007. 39

Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 27

Simone Undri Innocenti, Federico Becattini, Federico Pernici, and Alberto Del Bimbo. Temporal binary representation for event-based action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10426–10432. IEEE, 2021. 111, 114

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 36, 99, 118

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 50

Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *British Machine Vision Conference*, volume 2, page 5, 2018. 36, 37, 38, 39, 67, 74

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1): 221–231, 2012. 20, 23

Wenqi Jia, Miao Liu, and James M Rehg. Generative adversarial network for future hand segmentation from egocentric video. In *European Conference on Computer Vision*, pages 639–656. Springer, 2022. 26, 27

Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8484–8493, 2020. 40, 41

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 112

Xu Jiaolong, Xiao Liang, and Antonio M. López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019. 132, 137, 138, 142, 143

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43 (11):4037–4058, 2020. 41

Matthew Johnson and Yiannis Demiris. Perceptual perspective taking and action recognition. *International Journal of Advanced Robotic Systems*, 2(4):32, 2005. 19

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019. 36

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 69, 72

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019a. 57, 59, 69, 72

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019b. 20, 21, 27, 115

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021a. 59

Vangelis Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. In *The 32nd British Machine Vision Conference*, 2021b. 25, 26

Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics Symposium on Geometry Processing*, 2006. 152

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 64

Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, and Kris Kitani. Ego-humans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19807–19819, 2023. 150

Alireza Khodamoradi and Ryan Kastner. $o(n)$ o (n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. *IEEE Transactions on Emerging Topics in Computing*, 9(1):15–23, 2018. 48

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 39

Tarasha Khurana, Achal Dave, and Deva Ramanan. Detecting invisible people. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3174–3184, 2021. 150

Byungju Kim and Junmo Kim. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8:81674–81685, 2020. 80, 169

Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13618–13627, October 2021a. 28, 86

Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021b. 4, 72, 74, 109

Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J Davison. Simultaneous mosaicing and tracking with an event camera. *J. Solid State Circ*, 43:566–576, 2008. 48

Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2146–2156, 2021c. 51, 53, 111, 119

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 99

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 40

KM Kitani, T Okabe, Y Sato, and A Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3248, 2011. 19

Muhammed Kocabas, Ye Yuan, Pavlo Molchanov, Yunrong Guo, Michael J Black, Otmar Hilliges, Jan Kautz, and Umar Iqbal. Pace: Human and camera motion estimation from in-the-wild videos. *arXiv:2310.13768*, 2023. 150

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009. 53

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 22

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 36

Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016. 114

Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. In *33rd British Machine Vision Conference Proceedings, BMVC 2022*, 2022. 29

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the Association for Computational Linguistics*, 2023. 29

Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE international conference on robotics and automation*, pages 1817–1824. IEEE, 2011. 134

Matthias De Lange, Hamid Eghbalzadeh, Reuben Tan, Michael L. Iuzzolino, Franziska Meier, and Karl Ridgeway. Egoadapt: A multi-stream evaluation study of adaptation to real-world egocentric user video. *arXiv preprint arXiv:2307.05784*, 2023. 29

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 52

Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016. 48

Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012. 26

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017a. 82, 84, 86, 87

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a. 41

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, 2019a. 41

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018b. 40, 98, 100

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020. 40

Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017b. 53

Mengke Li, Yiu-Ming Cheung, and Juyong Jiang. Feature-balanced loss for long-tailed visual recognition. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2022a. 80, 169

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018c. 40, 41

Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017c. 36

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018d. 36, 72, 74

Yiming Li, Ziang Cao, Andrew Liang, Benjamin Liang, Luoyao Chen, Hang Zhao, and Chen Feng. Egocentric prediction of action target in 3d. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20971–20980. IEEE, 2022b. 31, 33

Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 287–295, 2015. 20

Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018e. 26

Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2019b. 41

Hui Liang, Junsong Yuan, Daniel Thalmann, and Nadia Magnenat Thalmann. Ar in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 743–744, 2015. 2, 19

Patrick Lichtsteiner et al. A 128 x 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 44

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019. 23, 59, 113, 114, 115, 118

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 28, 90

Yihan Lin, Wei Ding, Shaohua Qiang, Lei Deng, and Guoqi Li. Es-imagenet: A million event-stream classification dataset for spiking neural networks. *Frontiers in neuroscience*, 15:726582, 2021. 111

Bei Liu, S. Zheng, Jianlong Fu, and Wen-Huang Cheng. Anchor-based detection for natural language localization in ego-centric videos. In *International Conference on Consumer Electronics*, 2023. 29

Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020. 27

Miao Liu, Lingni Ma, Kiran Somasundaram, Yin Li, Kristen Grauman, James M Rehg, and Chao Li. Egocentric activity recognition and localization on a 3d map. In *European Conference on Computer Vision*, pages 621–638. Springer, 2022. 29, 32, 149

Mohammad Reza Loghmani, Luca Robbiano, Mirco Planamente, Kiru Park, Barbara Caputo, and Markus Vincze. Unsupervised domain adaptation through inter-modal rotation for rgb-d object recognition. *IEEE Robotics and Automation Letters*, 5(4):6631–6638, 2020. 125, 132, 134, 137, 141, 143

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 36, 72, 74, 131, 137, 138, 142, 143

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 19

Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2714–2721, 2013. 26

Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129:548–578, 2021. 159

Iulia-Alexandra Lungu, Federico Corradi, and Tobi Delbrück. Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–1. IEEE, 2017. xiv, 112

Jianming Lv, Kaijie Liu, and Shengfeng He. Differentiated learning for multi-modal domain adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, October 2021. 55

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2022. 29

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997a. 114

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997b. 47

Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, and Bernard Ghanem. Localizing objects in 3d from egocentric videos with visual queries. *arXiv preprint arXiv:2212.06969*, 2022. 31, 149, 153, 159

Sagnik Majumder, Hao Jiang, Pierre Moulon, Ethan Henderson, Paul Calamia, Kristen Grauman, and Vamsi Krishna Ithapu. Chat2map: Efficient scene mapping from multi-ego conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10554–10564, 2023. 31

Massimiliano Mancini, Zeynep Akata, Elisa Ricci, and Barbara Caputo. Towards recognizing unseen categories in unseen domains. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020. 41

Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5419–5427, 2018. 49, 125

Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023. 29

Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11749–11756, 2020. 40, 41

Georgios Meditskos, Pierre-Marie Plans, Thanos G Stavropoulos, Jenny Benois-Pineau, Vincent Buso, and Ioannis Kompatsiaris. Multi-modal activity recognition from egocentric vision, semantic enrichment and lifelogging applications for the care of dementia. *Journal of Visual Communication and Image Representation*, 51:169–190, 2018. 19

Boudjelal Meftah, Olivier Lezoray, and Abdelkader Benyettou. Segmentation and edge detection based on spiking neural network model. *Neural Processing Letters*, 32:131–146, 2010. 48

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Computer Vision and Pattern Recognition*, 2022. 150

Long Short-Term Memory. Long short-term memory. *Neural computation*, 9(8):1735–1780, 2010. 21

Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 86–104. Springer, 2020. 21

Shu Miao, Guang Chen, Xiangyu Ning, Yang Zi, Kejia Ren, Zhenshan Bing, and Alois Knoll. Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. *Frontiers in neurorobotics*, 13:38, 2019. xiv, 111, 112

Seonwoo Min, Nokyung Park, Siwon Kim, Seunghyun Park, and Jinkyu Kim. Grounding visual representations with texts for domain generalization. In *European Conference on Computer Vision*, pages 37–53. Springer, 2022. 42

M Keith Moore and Andrew N Meltzoff. Object permanence after a 24-hr delay and leaving the locale of disappearance: the role of memory, space, and identity. *Developmental Psychology*, 40(4):606, 2004. 147

Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744, 2020. 5

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 2

Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126:1381–1393, 2018. 48

Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a. xix, 4, 5, 28, 36, 38, 55, 67, 68, 70, 72, 74, 75, 82, 84, 86, 87, 109, 110, 111, 118

Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020b. 57, 59, 71

Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. *Advances in Neural Information Processing Systems*, 33: 2005–2015, 2020. 31

Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *Computer Vision and Pattern Recognition*, 2020. 149

Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egoenv: Human-centric environment representations from egocentric video. *Advances in Neural Information Processing Systems*, 36, 2024. 32, 149

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pages 892–909. PMLR, 2023. 29

Atsushi Nakazawa and Miwako Honda. First-person camera system to evaluate tender dementia-care skill. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 19

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 41

Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021. 24

Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. Actionflownet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1616–1624. IEEE, 2018. 55

Hongjing Niu, Hanting Li, Feng Zhao, and Bin Li. Domain-unified prompt representations for source-free domain generalization. *arXiv preprint arXiv:2209.14926*, 2022. 41

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 138

Peter O'Connor, Daniel Neil, Shih-Chii Liu, Tobi Delbruck, and Michael Pfeiffer. Real-time classification and sensor fusion with a spiking deep belief network. *Frontiers in neuroscience*, 7:178, 2013. 51

Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12999–13008, 2023. 28

Eshed OhnBar, Kris Kitani, and Chieko Asakawa. Personalized dynamics models for adaptive assistive navigation systems. In *Conference on Robot Learning*, pages 16–39. PMLR, 2018. 2

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 96

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 159, 162

Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015a. xvi, 134, 135, 136, 140

Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015b. 52, 53

Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, 2020. 36, 37, 38

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 4, 33, 149

Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4697–4705, 2016. 2

Razvan-George Pasca, Alexey Gavryushin, Yen-Ling Kuo, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction. *arXiv preprint arXiv:2301.09209*, 2023. 27, 29

Adam Paszke, S. Gross, Soumith Chintala, G. Chanan, E. Yang, Zachary Devito, Zeming Lin, Alban Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017. 118

Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*, 2020. 87

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in neural information processing systems*, 34:12493–12506, 2021. 20, 24, 25

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 86, 87

José Antonio Pérez-Carrasco, Bo Zhao, Carmen Serrano, Begona Acha, Teresa Serrano-Gotarredona, Shouchun Chen, and Bernabé Linares-Barranco. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward convnets. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2706–2719, 2013. 48, 52

Etienne Perot, Pierre De Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020. 111

Toby Perrett and Dima Damen. Ddlstm: dual-domain lstm for cross-dataset action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7852–7861, 2019. 21

Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 475–484, 2021. 87

Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Use your head: Improving long-tail video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2415–2425, 2023. 87

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021. 120

Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 19, 26

Mirco Planamente, Chiara Plizzari, Marco Cannici, Marco Ciccone, Francesco Strada, Andrea Bottino, Matteo Matteucci, and Barbara Caputo. Da4event: towards bridging the sim-to-real gap for event cameras using domain adaptation. *IEEE Robotics and Automation Letters*, 6(4):6616–6623, 2021. 8, 114, 123

Mirco Planamente, Gabriele Goletto, Gabriele Trivigno, Giuseppe Averta, and Barbara Caputo. Polito-iit-cini submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition. *arXiv preprint arXiv:2209.04525*, 2022a. 68

Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1807–1818, 2022b. 7, 69, 82, 84, 109

Mirco Planamente, Chiara Plizzari, Simone Alberto Peirone, Barbara Caputo, and Andrea Bottino. Relative norm alignment for tackling domain shift in deep multi-modal classification. *International Journal of Computer Vision*, pages 1–21, 2024. 7

Chiara Plizzari, Mirco Planamente, Emanuele Alberti, and Barbara Caputo. Polito-iit submission to the epic-kitchens-100 unsupervised domain adaptation challenge for action recognition. *arXiv preprint arXiv:2107.00337*, 2021. 68

Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, and Barbara Caputo. E2 (go) motion: Motion augmented event stream for egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19935–19947, 2022. 8

Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *International Journal of Computer Vision*, 2023a. 7, 12

Chiara Plizzari, Toby Perrett, Barbara Caputo, and Dima Damen. What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13656–13666, 2023b. 7, 29

Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. *arXiv preprint arXiv:2404.05072*, 2024. 8

Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. Compact cnn for indexing egocentric videos. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 20

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 2018. 57, 59

Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits*, 46(1):259–275, 2010. 45

Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: bioinspired cameras with spiking output. *Proceedings of the IEEE*, 102(10):1470–1484, 2014. xi, 45

Will Price and Dima Damen. An evaluation of action recognition models on epic-kitchens. *arXiv preprint arXiv:1908.00867*, 2019. 113

Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. ISSN 0893-6080. 118

Shengyi Qian and David F Fouhey. Understanding 3d object interaction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21753–21763, 2023. 32, 33

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12556–12565, 2020. 41

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 103, 162

Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023. 29

Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021. 28

Francesco Ragusa, Giovanni Maria Farinella, and Antonino Furnari. Stillfast: An end-to-end approach for short-term object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3644, 2023a. 29

Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. *Computer Vision and Image Understanding*, 235:103764, 2023b. 26, 28

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Correlation-aware adversarial domain adaptation and generalization. *Pattern Recognition*, 100:107124, 2020. 40

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. In *Advances in Neural Information Processing Systems*, 2021. 150

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3d appearance, location and pose. In *Computer Vision and Pattern Recognition*, 2022. 150, 155

Santhosh Kumar Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6694–6703, 2023. 29

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 29

Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. L2-constrained softmax loss for discriminative face verification. *arXiv preprint arXiv:1703.09507*, 2017. 60

Nitin Rathi, Priyadarshini Panda, and Kaushik Roy. Stdp-based pruning of connections and weight quantization in spiking neural networks for energy-efficient recognition. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 38(4):668–677, 2018. 48

Siddharth Ravi, Pau Climent-Perez, Théo Morales, Carlo Huesca-Spairani, Kooshan Hashemifard, and Francisco Florez-Revuelta. Odin: An omnidirectional indoor dataset capturing activities of daily living from multiple synchronized modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6487–6496, 2023. 34, 149

Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 16–1, 2017. 49

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on robot learning*, pages 969–982. PMLR, 2018. xxi, 54, 112, 125, 127, 128, 135, 136

Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981, 2013. 53

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. Association for Computational Linguistics, 2019. 99

Christian Reinbacher, Gottfried Munda, and Thomas Pock. Real-time panoramic tracking for event cameras. In *2017 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2017. 48

Xuanchi Ren, Tao Yang, Li Erran Li, Alexandre Alahi, and Qifeng Chen. Safety-aware motion prediction with unseen vehicles for autonomous driving. In *International Conference on Computer Vision*, 2021. 150

Nicholas Rhinehart and Kris M Kitani. Learning action maps of large environments via first-person vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–588, 2016. 31

Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*. Springer, 2016. 159

Debaditya Roy and Basura Fernando. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2745–2753, 2022. 27

Bodo Rueckauer and Tobi Delbruck. Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor. *Frontiers in neuroscience*, 10:176, 2016. 51

Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017. 48

Michael S Ryoo and Larry Matthies. First-person activity recognition: What are they doing to me? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2730–2737, 2013. 19

Michael S Ryoo, Brandon Rothrock, and Larry Matthies. Pooled motion features for first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 896–904, 2015. 20

Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *Advances in Neural Information Processing Systems*, 34:23386–23400, 2021a. 28

Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021b. 38, 39, 109

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018. 36, 72, 74

Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 134

Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer, 2018. 48

Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Asynchronous spatial image convolutions for event cameras. *IEEE Robotics and Automation Letters*, 4(2):816–822, 2019. 48

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition*, 2016. 151

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. 152

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. xvi, 139, 141

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 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*, pages 21096–21106, 2022. 26, 28

Teresa Serrano-Gotarredona and Bernabé Linares-Barranco. A 128\,×128 1.5% contrast sensitivity 0.9% fpn 3 μs latency 4 mw asynchronous frame-free dynamic vision sensor using transimpedance preamplifiers. *IEEE Journal of Solid-State Circuits*, 48(3):827–838, 2013. 44

Teresa Serrano-Gotarredona and Bernabé Linares-Barranco. Poker-dvs and mnist-dvs. their history, how they were made, and other details. *Frontiers in neuroscience*, 9:481, 2015. 52

Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. In *German conference on pattern recognition*, pages 281–297. Springer, 2019. 3, 55

Aviv Shamsian, Ofri Kleinfeld, Amir Globerson, and Gal Chechik. Learning object permanence from video. In *European Conference on Computer Vision*, 2020. 149

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020. xviii, 157, 162, 164

Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10031, 2019. 41

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR, 2023. 28

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 41

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 22

Suriya Singh, Chetan Arora, and CV Jawahar. First person action recognition using deep learned descriptors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2620–2628, 2016. 20

Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018a. 114, 133

Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1731–1740, 2018b. 50, 53

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 103

Min-Ho Song and Rolf Inge Godøy. How fast is your body motion? determining a sufficient frame rate for an optical motion tracking system using passive markers. *PloS one*, 11(3): e0150993, 2016. 122

Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9787–9795, June 2021a. 86

Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9787–9795, 2021b. 36, 38, 39, 72, 74, 109

Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 17–24. IEEE, 2009. 19

Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7244–7253, 2019. 144, 171

Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 534–549. Springer, 2020. 114, 123, 125, 127, 139, 140, 143

Swathikiran Sudhakaran and Oswald Lanz. Convolutional long short-term memory networks for recognizing first person interactions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2339–2346, 2017. 20, 21

Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *29th British Machine Vision Conference, BMVC 2018; Proceedings*. BMVA Press, 2018. 20, 21

Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9954–9963, 2019. 21

Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1111, 2020. 23

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016. 36, 40, 98

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018a. 122

Lin Sun, Kui Jia, Kevin Chen, Dit-Yan Yeung, Bertram E Shi, and Silvio Savarese. Lattice long short-term memory for human action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2147–2156, 2017. 21

Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018b. 55

Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. Trace: 5d temporal regression of avatars with dynamic cameras in 3d environments. In *Computer Vision and Pattern Recognition*, 2023. 150

Dipak Surie, Thomas Pederson, Fabien Lagriffoul, Lars-Erik Janlert, and Daniel Sjölie. Activity recognition using an egocentric perspective of everyday objects. In *Ubiquitous Intelligence and Computing: 4th International Conference, UIC 2007, Hong Kong, China, July 11-13, 2007. Proceedings 4*, pages 246–257. Springer, 2007. 19

Dídac Surís, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. Learning to learn words from visual scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 434–452. Springer, 2020. 28

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 41

Hao Tang, Kevin J Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 29, 149

Hui Tang and Kui Jia. Discriminative adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5940–5947, 2020. 36

Yansong Tang, Yi Tian, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Action recognition in rgb-d egocentric videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3410–3414. IEEE, 2017. 20

Koya Tango, Takehiko Ohkawa, Ryosuke Furuta, and Yoichi Sato. Background mixup data augmentation for hand and object-in-contact detection. In *ECCV Workshop*, 2022. 28

Catherine Taylor, Robin McNicholas, and Darren Cosker. Towards an egocentric framework for rigid and articulated object tracking in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 354–359. IEEE, 2020. 2, 19

Daksh Thapar, Aditya Nigam, and Chetan Arora. Recognizing camera wearer from hand gestures in egocentric videos. In *International Conference on Multimedia*, 2020. 27

Pavel Tokmakov, Jie Li, Wolfram Burgard, and Adrien Gaidon. Learning to track with object permanence. In *International Conference on Computer Vision*, 2021. 149

A Torralba and AA Efros. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011. 2, 57, 82, 84, 86, 87

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 20, 23

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 20, 23

Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36, 2024. x, 5, 27, 33, 34, 151, 157

Pavan Turaga, Ashok Veeraraghavan, and Rama Chellappa. Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 39

Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. In *Computer Vision and Pattern Recognition*, 2023. 149

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE trans. neural netw.*, 10 (5):988–999, 1999. 98

Ajay Vasudevan, Pablo Negri, Bernabe Linares-Barranco, and Teresa Serrano-Gotarredona. Introduction and analysis of an event-based sign language dataset. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 675–682. IEEE, 2020. xiv, 112

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 20, 23, 24, 94

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 86, 87

Sagar Verma, Pravin Nagar, Divam Gupta, and Chetan Arora. Making third person techniques recognize first-person actions in egocentric videos. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2301–2305. IEEE, 2018. 20

Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989, 2019. 41

Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *Advances in neural information processing systems*, pages 5334–5344, 2018. 40, 41

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016. 26

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7364–7373, 2019. 134

David Ed Waller and Lynn Ed Nadel. *Handbook of spatial cognition.* American Psychological Association, 2013. 147, 159

Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017a. 60

Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5250–5261, 2023a. 29

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 20, 59, 72, 118

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4325–4334, 2017b. 21

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 113, 114, 118

Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1895–1904, 2021a. 23

Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10081–10090, 2019a. 49

Tian Wang and Hichem Snoussi. Histograms of optical flow orientation for abnormal events detection. In *2013 IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pages 45–52. IEEE, 2013. 19

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020a. 55, 57, 59, 69, 70, 71

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020b. 103

Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:5345–5352, July 2019b. 67

Xizi Wang, Feng Cheng, Gedas Bertasius, and David J. Crandall. Loconet: Long-short context network for active speaker detection. *arXiv preprint arXiv:2301.08237*, 2023b. 29

Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3622–3626. IEEE, 2020c. 41, 99

Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13214–13223, 2021b. 23

Ziqi Wang, Marco Loog, and Jan Van Gemert. Respecting domain relations: Hypothesis invariance for domain generalization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9756–9763. IEEE, 2021c. 40

Pengfei Wei, Lingdong Kong, Xinghua Qu, Xiang Yin, Zhiqiang Xu, Jing Jiang, and Zejun Ma. Unsupervised video domain adaptation: A disentanglement perspective. *arXiv preprint arXiv:2208.07365*, 2022. 4, 38, 40, 67

David Weikersdorfer, David B Adrian, Daniel Cremers, and Jörg Conradt. Event-based 3d slam with a depth-augmented dynamic vision sensor. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 359–364. IEEE, 2014. 51

Yuhang Wen, Zixuan Tang, Yunsheng Pang, Beichen Ding, and Mengyuan Liu. Interactive spatiotemporal token attention network for skeleton-based general interactive action recognition. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7886–7892. IEEE, 2023. 28

Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011. 57, 59

Kelvin Wong, Yanlei Gu, and Shunsuke Kamijo. Mapping for autonomous driving: Opportunities and challenges. *IEEE Intelligent Transportation Systems Magazine*, 13(1):91–106, 2020. 150

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022a. 25

Junru Wu, Yi Liang, Hassan Akbari, Zhangyang Wang, Cong Yu, et al. Scaling multimodal pre-training via cross-modality gradient harmonization. *Advances in Neural Information Processing Systems*, 35:36161–36173, 2022b. 169

QingXiang Wu, Martin McGinnity, Liam Maguire, Ammar Belatreche, and Brendan Glackin. Edge detection based on spiking neural network model. In *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence: Third International Conference on Intelligent Computing, Proceedings 3*, pages 26–34. Springer, 2007. 48

Yue Wu, Hongfu Liu, Jun Li, and Yun Fu. Deep face recognition with center invariant loss. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 408–414, 2017. 80, 169

Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&m mix: A multimodal multiview transformer ensemble. *arXiv preprint arXiv:2206.09852*, 2022. 27

Mengmeng Xu, Yanghao Li, Cheng-Yang Fu, Bernard Ghanem, Tao Xiang, and Juan-Manuel Pérez-Rúa. Where is my wallet? modeling object proposal sets for egocentric visual query localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2593–2603, 2023. 31

Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6502–6509, 2020. 41

R. Xu, G. Li, J. Yang, and L. Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1426–1435, 2019a. 36, 58, 65

Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019b. 143

Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1435, 2019c. 132, 137, 138, 142

Yuecong Xu, Haozhi Cao, Zhenghua Chen, Xiaoli Li, Lihua Xie, and Jianfei Yang. Video unsupervised domain adaptation with deep learning: A comprehensive survey. *arXiv preprint arXiv:2211.10412*, 2022. 37

Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022. 25, 27

Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 480–498. Springer, 2020. 39

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 152

Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14722–14732, 2022a. 38, 68, 72, 74

Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, and Du Tran. Relational space-time query in long-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6408, 2023. 29

Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 158

Yuzhe Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *European Conference on Computer Vision*, pages 57–75. Springer, 2022b. 40, 99

Zhiyu Yao, Yunbo Wang, Jianmin Wang, Philip Yu, and Mingsheng Long. Videodg: Generalizing temporal relations in videos to novel domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 42

Jianbo Ye, Xin Lu, Zhe Lin, and James Z Wang. Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 60

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Computer Vision and Pattern Recognition*, 2023. 150

Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. Glamr: Global occlusion-aware human mesh recovery with dynamic cameras. In *Computer Vision and Pattern Recognition*, 2022. 150

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 22, 107, 109, 122

Olga Zatsarynna and Juergen Gall. Action anticipation with goal consistency. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1630–1634. IEEE, 2023. 28

Jeroen Zewald and Ivo Jacobs. Object permanence. In *Encyclopedia of animal cognition and behavior*, pages 4711–4727. Springer, 2022. 147

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 120

Kai Zhan, Steven Faux, and Fabio Ramos. Multi-scale conditional random fields for first-person activity recognition. In *2014 IEEE international conference on pervasive computing and communications (PerCom)*, pages 51–59. IEEE, 2014. 19

Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2718–2726, 2016. 22

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 41

Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022a. 29

Xin Zhang, Yusuke Iwasawa, Yutaka Matsuo, and Shixiang Shane Gu. Amortized prompt: Lightweight fine-tuning for clip in domain generalization. *arXiv preprint arXiv:2111.12853*, 2021. 41

Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8035–8045, 2022b. 41

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*. Springer, 2022c. 150

Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees GM Snoek. Audio-adaptive activity recognition across video domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13791–13800, 2022d. 38, 39

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 28, 30

Xiaozheng Zheng, Chao Wen, Zhou Xue, and Jingyu Wang. Hand pose estimation via multiview collaborative self-supervised learning. *arXiv preprint arXiv:2302.00988*, 2023. 28

Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5089–5097, 2018. 60

Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt vision transformer for domain generalization. *arXiv preprint arXiv:2208.08914*, 2022. 41, 42, 99

Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023. 26, 27

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 21, 37, 69, 99

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4396–4415, 2022. 41, 84, 86

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Mixstyle neural networks for domain generalization and adaptation. *International Journal of Computer Vision*, pages 1–15, 2023. 41

Alex Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, June 2018. doi: 10.15607/rss.2018.xiv.062. 49, 51

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019a. xv, xvi, 112, 114, 128, 130, 133, 135, 141

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019b. 49, 50

Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18(3):351–376, 2021. 55

Ming Zong, Ruili Wang, Xiubo Chen, Zhe Chen, and Yuanhao Gong. Motion saliency based multi-stream multiplier resnets for action recognition. *Image and Vision Computing*, 107: 104108, 2021. 22