UNIVERSITY OF CALIFORNIA
RIVERSIDE

# Exocentric To Egocentric Transfer For Action Recognition

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Computer Science

by

Anirudh Thatipelli

September 2024

Thesis Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Greg Ver Steeg
Dr. Emiliano De Cristofaro

The Thesis of Anirudh Thatipelli is approved:

Committee Chairperson

University of California, Riverside

## Acknowledgments

I would like to extend my deepest gratitude to my thesis advisor, Dr. Amit Roy-Chowdhury. His mentorship and unwavering support was monumental for the completion of this thesis. His constant feedback has been invaluable to my research journey.

I would like to thank Dr. Greg and Dr. Emiliano for graciously accepting to be a part of my committee. Their teaching prowess expanded my horizon. I thoroughly enjoyed the challenging questions and thought-provoking discussions that enhanced my own perspective.

I am also thankful to Prof Naell and Prof Yue Dong for their valuable discussions.

I am grateful to my family: my mother Padma Priya Thatipelli, my father Sathish Kumar Thatipelli and my sister Anusha Thatipelli for their constant support.

I feel blessed to be surrounded by an outstanding cohort of students at UCR. I am thankful to my roommates, Arpit Mallick, Sahil Chowkekar and Priyanshu Sharma for the wonderful memories. I cherish the time spent with Puneet, Manoj, Abhav, Gayatri, Rohit, Saketh, Yash and Vineeth. Thanks Rinki and Sayak for challenging discussions.

I am especially grateful to my friend and collaborator, Erfan for his constructive criticism and insights for my research.

My collaborator Dr. Shao-Yuan gave valuable input and suggestions for this project.

A special acknowledgement to Victor Hill for his patience and maintaining the cluster enabling my research.

I feel priviliged to have been mentored by: Dr. Sanath Narayan, Dr. Fahad Khan, Dr. Salman Khan, Dr. Ravi Kiran Sarvadevabhatla.

**Acknowledgement of previously published materials**. The text of this thesis, in part or in full, is a reprint of published/under-review material.

To my family for all the support.

# ABSTRACT OF THE THESIS

Exocentric To Egocentric Transfer For Action Recognition

by

Anirudh Thatipelli

Master of Science, Graduate Program in Computer Science
University of California, Riverside, September 2024
Dr. Amit K. Roy-Chowdhury, Chairperson

Egocentric vision captures the scene from the point of the view of the camera wearer while exocentric vision captures the overall scene context. Jointly modelling ego and exo views is a crucial step towards developing next-generation AI agents. The community has regained interest in the field of egocentric vision. While, third-person view and first-person has been thoroughly investigated, very few works aim to study the both synchronously. Exocentric videos contain many relevant signals transferrable to egocentric videos. We propose a multimodal-LLM model that leverages large-scale exocentric information for the task of egocentric action recognition. This thesis also provides a broad overview of works combining both the egocentric and exocentric vision.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Human beings perceive the world from multiple viewpoints. We watch Do-it-yourself videos to learn new skills. A bicycle repair video alternates between the ego (1st-person) and exo (3rd-person) viewpoints. An ego (close-up) view of the bicycle captures vital hand-object interactions and an exo (third-person) view captures the overall context in the environment. We are able to relate the object from 3rd-person to 1st person perspective. Being able to map skills to one's own body has been a well-studied problem in cognitive science [40, 94, 119]. Capturing video from both the **ego** and **exo** views is a vital frontier for AI to understand human activities. Widespread applications exist in augmented reality [113, 95] and robotics [64, 92, 116].

Despite the importance of multi-view learning, most efforts into video understanding have focused to only one view, 3rd-person (exocentric) viewpoint [127, 35, 20, 3, 4, 71] or 1st-person (egocentric) viewpoints [43] separately. While existing algorithms perform considerably well on 3rd-person settings [1], a significant gap exists in the egocentric

Figure 1.1: Hand-object interactions in the 3rd person view(right) are useful for identifying the action from the 1st person viewpoint(left).

settings [44, 24, 23].

Exocentric view contains many relevant cues for recognizing in the egocentric view. For example in Fig 1.1, the hand-object interaction of ”cutting” in $3^{rd}$-person view can be useful to recognize in the $1^{st}$-person view.

Existing vision-language models [107, 61] trained on large amounts of 3rd person perspective contain many signals that can be transferable for egocentric tasks. Previous works like [75, 11] utilize simpler architectures to learn egocentric representations from 3rd-person data. VLMs are more capable of learning stronger representations.

The intention of this thesis is twofold:

- Propose an approach that leverages the exocentric signals embedded in VLMs for solving egocentric vision task.
- Provide a high-level overview of the various egocentric-exocentric learning tasks.

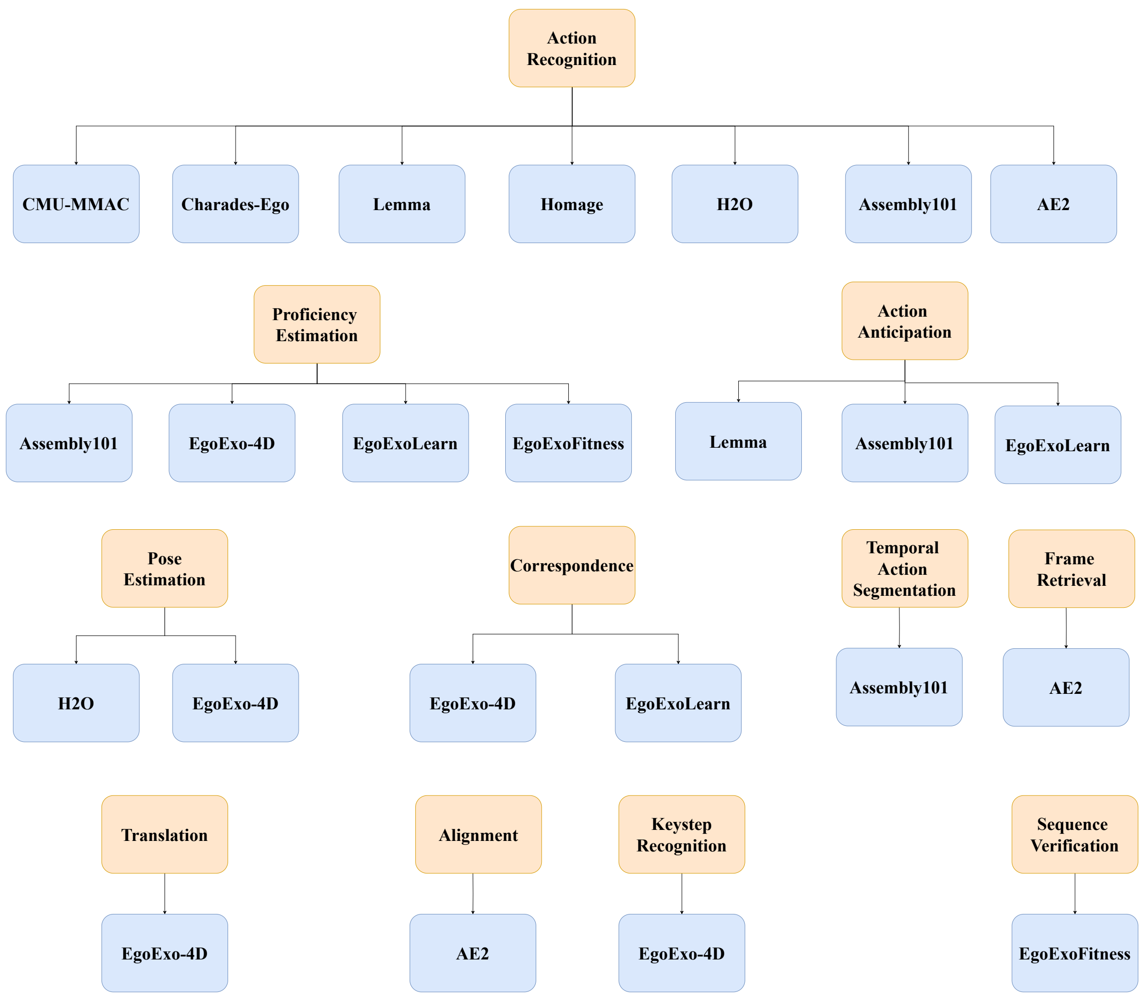


Figure 1.2: **Ego-Exo Datasets and corresponding tasks.** This figure illustrates the different Ego-Exo datasets in literature and compares them with respect to the associated benchmarks. Newly released Ego-Exo4D [44], EgoExoLearn [57], EgoExoFitness [76] constitute a large suite of novel tasks to further research in this arena.

# Chapter 2

# Datasets

Several datasets containing paired ego-exo views have been proposed in the literature [69, 70, 110, 122, 125]. ”Mixed” ego-exo views have been covered by [89, 133, 160, 161, 146, 19, 22, 67, 145]. Zhang *et al.* [154] captures egocentric interactions in a 3D viewpoint. Xue *et al.*'s AE2 dataset [145] is sampled from multiple existing ego and exo datasets. These datasets have several shortcomings: lack of magnitude, weak synchronization and poor diversity. The release of two new large-scale datasets [44, 57, 76] attempts to bridge this gap. Refer Table 2.1 for an in-depth review.

The **CMU-MMAC dataset** [70] is one of the earliest dataset that captures ego and exo video. It is composed of 43 participants cooking 5 recipes in the kitchen setting. Multiple modalities like audio, video, accelerations, and motion capture are present in this dataset.

The **Charades-Ego dataset** [125] was one of the former large-scale joint multi-view dataset efforts containing 68.8 hours of first and third-person video. 112 actors hired

Table 2.1: Comparison of existing datasets across various parameters, arranged in a chronological order.

| Dataset | Year Published | Hours | Num. action clips | Num. scenarios | Num. subjects | Num. verb classes | Num. noun classes | Num. action classes |
|---|---|---|---|---|---|---|---|---|
| CMU-MMAC [70] | 2009 | - | 5 | 1 | 43 | - | - | 8 |
| Charades-Ego [125] | 2018 | 68.8 | 68536 | 15 | 112 | 33 | 36 | 157 |
| Lemma [60] | 2020 | 10.1 | 324 | 15 | 8 | 24 | 64 | 641 |
| Homage [110] | 2021 | 25.4 | 453 | 70 | 27 | 29 | 86 | - |
| H2o [69] | 2021 | 5 | 500 | 36 | 4 | 11 | 8 | 36 |
| Assembly101 [122] | 2022 | 513 | 4321 | 15 | 53 | 24 | 90 | 1380 |
| AE2 [145] | 2023 | - | 322 | 6 | - | - | - | 4 |
| Ego-Exo4D [44] | 2024 | 1286 | 5035 | 43 | 740 | 1481 | 2924 | 689 |
| EgoExoLearn [57] | 2024 | 120 | 747 | 8 | - | 95 | 254 | 39 |
| EgoExo-Fitness [76] | 2024 | 31 | 1248 | 1 | 49 | - | - | 76 |

by Amazon Mechanical Turk recorded 34 hours of scripted scenarios.

The **Lemma dataset** [60] is composed of multi-view and multi-agent daily-life activities. 3D skeletons and RGB-D are collected to give a broad perspective.

The **Homage dataset** [110] is a synchronized multi-view dataset consisting of 30 hours of ego-exo video from 27 participants performing household activities in the same environment. It is well annotated with both the hierarchical and atomic action-labels.

The **H2O dataset** [69] focuses on 3D egocentric object-level manipulations. It is composed of 3D hand-poses, 6D object poses, camera poses, object meshes and scene-point clouds. 4 different participants perform 36 unique actions in three unique environments.

The **Assembly101 dataset** [122] features non-scripted multi-step activities. 101 toy-vehicles are manipulated in 4321 video sequences for a total of 513 hours. It constitutes

1380 fine-grained and 202 coarse-grained action classes. **AssemblyHands** [99] is a subset of Assembly101 to study the challenging problem of 3D hand pose estimation and action classification.

The **AE2 dataset** [145] is one of the premier attempts to learn a view-invariant self-supervised embedding from unpaired ego and exo videos. To this end, they create a new benchmark, sampled from five public datasets [23, 70, 68, 69, 156], and a self-collected tennis dataset. It is composed of 322 clips.

The **Ego-Exo4D dataset** [44] is the largest multi-view dataset including egocentric view and the corresponding exocentric information. Moreover, it also offers multiple natural language descriptions including expert commentary, narrate-and-act descriptions and atomic action descriptions. It is rich in modalities like audio, IMU, video, depth, gaze, stereo, 3D environments, thermal IR, GPS, motion capture, 6DOF, barometer and magnetometer readings. 740 subjects shot 123 scenes across different cities. It releases new challenging benchmarks like keystep recognition, efficient action detection and proficiency estimation.

The **EgoExoLearn dataset** [57] is another concurrent large-scale ego-exo synchronized dataset. It contains 120 hours of demonstration activities recorded in the lab and daily-life settings. It is richly annotated with fine-grained captions. Unlike previous datasets, it releases benchmarks on cross-view action anticipation and proficiency estimation.

The **EgoExo-Fitness dataset** [76] was also concurrently released along with the previous two ego-exo datasets. While the previous datasets extensively explored daily-life activities, EgoExo-Fitness focuses exercise-related activities. It comes with a new set of benchmarks for cross-view sequence verification.

# Chapter 3

# Related work

Egocentric vision focuses on camera-wearer centric cues while exocentric vision focuses on a broader perspective of the subject in the context of the entire scene. Leveraging the complementary signals from both the viewpoints will enable us to learn human skill effectively.

Some early work has investigated the task of jointly relating egocentric and exocentric vision [5, 128]. In this section, we discuss important tasks jointly modelling vision from first-person and third-person perspective.

### 3.0.1 Identification

It is the task of matching a camera wearers in a egocentric video to an exocentric video. Lack of visibility in the egocentric video makes this task challenging. Being able to match a participant in both views is an important preliminary task for joint ego-exo learning. It is a well-researched problem. Ardeshir *et al.*[6, 7] is one of the first works that proposes a graph-matching technique to solve this problem. Ardeshir *et al.*[9, 7] further propose a joint approach to tackle temporal alignment and person re-identification. Fig 3.2 shows some examples of this dataset. Similarly, Han *et al.* [50] propose a different matching function based on spatial distributions. Fan *et al.* [36] learns a joint-embedding space. The model proposed by Ardeshir *et al.* [10] extended to focus on temporal alignment. In Han *et al.* [51], a conditional random field is proposed to identify the subjects in different viewpoints. Xu *et al.* [144] perform simultaneous matching and segmentation of the subject across both the views.

While identification focuses on solely matching the camera wearer, re-identification aims to learn the associations between the different subjects present in the egocentric and exocentric views. Work by Ardeshir *et al.* [12] is one of the earliest approaches exploring the task of re-identification between the different views. To enable further research in multi-view video-based re-identification, Basaran *et al.*[18] release a novel multimodal dataset, consisting of around $176,000$ detections. Han *et al.* [50] utilizes the spatial information such as the view-angle of the camera to perform the association. Han *et al.* [48] attempts to solve a challenging version of the problem by assuming limited appearance matching and different viewing angles in the ego and exo image. Example images from this dataset can be seen in 3.3.Han *et al.*[49] considers another challenging variant of this matching problem having minimal overlap of the field-of-view.

### 3.0.2 Action Recognition

Action Recognition is the task of identifying or assigning a category or multiple categories to the action performed by the subject in the video. The release of GoPro wearable cameras led to a large production of first-person videos. However, limited works have combined ego and exo views for identifying actions. The earliest attempt to recognize human activity across first and multiple third-person cameras was done in [128]. It presents a learnable weighted importance classification approach. Truong *et al.* [136] learns a geometric constraint to transfer knowledge between the multiple views. Rocha *et. al* [120] learns an invariant space, based on skeleton pose information. Huang *et al.* [56] extends to a multi-domain scenario, learning a holographic feature space based on both view-invariant and view-specific features. In Peng *et al.*[101], virtual features from first-person perspective are synthesisized and combined to perform action recognition. Different from other works, Ramirez *et al.* [111] incorporates gaze information into the robot's internal representation for improved imitation of human behaviour.

### 3.0.3 Tracking

Tracking is an important Computer Vision problem, where we estimate the global trajectories and match subjects across the video. Yang *et al.* [148] is one of the earliest works that jointly identifies and tracks the subjects across the first and third-person views. A deep neural network (DNN), robust to action and motion changes is used to generate the 3D trajectory. Han *et al.* [47] learns a spatio-temporal correspondence between the images of different viewpoints. 3.4 shows some sample frames released in their dataset [47].

In a follow-up work, Han *et al.* [?] treats the task as a joint optimization problem. Han *et al.*[45] extends the optimization approach for relating a single third-person view with multiple first-person view images. Recent work by DivoTrack [52] presents a new baseline for multi-view object tracking. Multi-view tracking has also gained importance in other areas like robotics [78].

### 3.0.4 Generation

In generation, we aim to synthesize an egocentric image, conditioned on an exocentric image and vice versa. Elfeki *et al.* [34] was the first landmark dataset for exo-ego synthesis and retrieval. A conditional GAN [91] is used to syntheize first-person images. Refer to Fig. 3.5 for example frames. Liu *et al.* [80] also utilize a variation of a GAN. Similarly, Tang *et al.* [132, 131] utilize semantic information to generate images in different views. Liu *et al.* [79] utilize a shared network between the ego-exo frames to aid generation. Liu *et al.* [81] syntheize egocentric videos by combining the semantic map with GANs. Recent work by Luo *et al.* [87] presents a diffusion-based technique [54] for exocentric to egocentric video synthesis. Different from all the other works, Luo *et al.* [86] uses action description and egocentric frames to synthesize a video from the third-person perspective. The new Ego-Exo4D dataset [44] constitutes a benchmark for synthesis.

A lot of progress has been made in a related problem of aerial view to ground view synthesis [117, 134, 130].

### 3.0.5 Affordance

Much attention has been drawn to affordance [41, 63, 14]. The objective is to understand the different possible actions that can be performed with an object. Luo *et al.* [84, 85] extracts affordance-level features from exocentric human-object interactions and transfers it to the egocentric view. Li *et. al* [73] extend the same work, but use a weakly-supervised technique. Chen *et al.* [21] extends affordance learning from videos using an attention-based network. Xu *et al.* [**?**] also uses a weakly-supervised technique leveraging cross-view knowledge. Recent work by Zhang *et al.* [157] integrates a self-explainable module to aid affordance learning. Yang *et al.* [147] presents a joint coarse and fine-grained feature extraction technique. Different from other techniques, Rai *et al.* [109] leverage VLM's knowledge as an auxiliary mask for the task of grounding. Check Fig. 3.6 for images and corresponding affordance.

### 3.0.6 Exo-Ego Transfer

A vast amount of knowledge in the form of motion cues is embedded in exocentric videos that can be transferred to the egocentric domain. Ardeshir *et al.* [11] is a premier work that learns mappings between the ego-exo views. In Ardeshir *et al.* [8], the authors propose a two-stream view-specific architecture to adapt from exo to ego view. Ho *et al.* [53] utilizes a semi-supervised domain adaptation technique to adapt exocentric visual cues to egocentric videos. Xu *et al.* [141] uses a prompt-masking technique for transferring information for egocentric hand-object interaction. Different from previous approaches, Li *et al.* [75] proposes an improved pre-training approach to extract signals from exocentric videos helpful for the egocentric domain. Ohkawa *et al.* [100] aids further adaptation by performing view-invariant pretraining and finetuning. Different from previous techniques, Quattrocchi *et al.* [106] proposes an adaptation technique for temporal action segmentation.

In Nishimura *et al.* [97], geometric transformation is used to tackle a novel problem of view-birdification(bird's eye-view trajectory estimation) is computed from the egocentric movement. Qian *et al.* [105] is an extension to a more challenging problem of bird's eye view estimation in the absence of proper calibration.

### 3.0.7 Joint ego-exo works

This section outlines works that aim to learn a joint ego-exo representation. Sigurdsson *et al.* [124] makes the first attempt to jointly relate first-person and third-person viewpoints. Yu *et al.* [150, 151] leverages a joint attention mechanism to extract a shared representation between the views. In Wang *et al.* [138], a sentence-bert language model [118] is utilized to semantically align the unpaired exocentric and egocentric videos. Xue *et al.*[145] became the first work to propose a self-supervised learning approach to learn a view-invariant representation. Zhao *et al.* [159] solves a novel task of identifying and segmenting the egocentric camera wearer in a third-person view.

3D egocentric pose estimation has also benefited from a joint ego-exo learning framework [27, 82]. A novel thermal image-based 3D hand-pose dataset has been released in ThermoHands [29]. Lu *et al.* [83] covers a scene-graph generation technique based on a self-attention mechanism between the ego and exo views. The authors of Wen *et al.* [139] present a solution combining 3rd person and 1st person images to predict the subject's location in the 3rd person viewpoint. Jia *et al.* [62] extracts exocentric and egocentric conversational signals to generate a scene-graph. Xu *et al.* [142] shows an improvement in egocentric captioning by retrieving semantically relevant 3rd person videos [15].

### 3.0.8 Miscellaneous applications

Jointly relating exocentric and egocentric vision has applications in Robotics and Virtual Reality. Kennedy *et al.* [66] illustrate the importance of combining egocentric and exocentric information for mapping. Multi-view visual feedback to the robots of the swarm can improve performance [115]. This has been corroborated in robotics manipulation as well [59]. Supervision from third-person videos have been well-adapted to egocentric vision in robotics [123, 17, 129]. A combination of hand and third-person perspective has been used in [55]. Young *et al.* [149] demonstrates the superior performance on the aerial telemanipulation task using egocentric-exocentric views. Abdullash *et al.* [2] synthesize third-person view from first-person view for enhanced teleoperation. Video captioning has also benefited from a joint ego-exo information [64].

Combining ego and exo views has been thoroughly researched in virtual reality [25, 90, 31]. Multiple works [42, 112, 152] demonstrate the possibility of using a mixed viewpoint space for collaboration. Soares *et al.* [126] proposes a novel cooperative virtual environment with fixed freedom of movement per user. Peschel *et al.* [102] illustrates the use-case of a joint ego-exo system for unmanned aerial systems. Automatic rendevous and docking (ARD) also benefits from a joint view system [72]. Duncan *et al.* [30]'s work proposes a camera system to reconstruct embodied experiences in real-time.

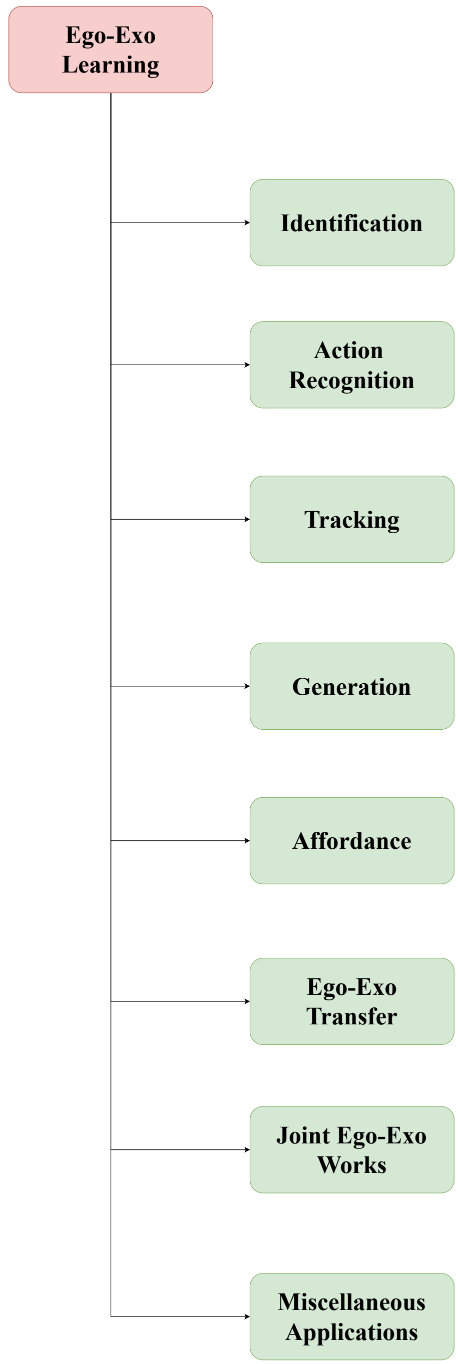


Figure 3.1: Overview of the landscape of various joint egocentric and exocentric applications.

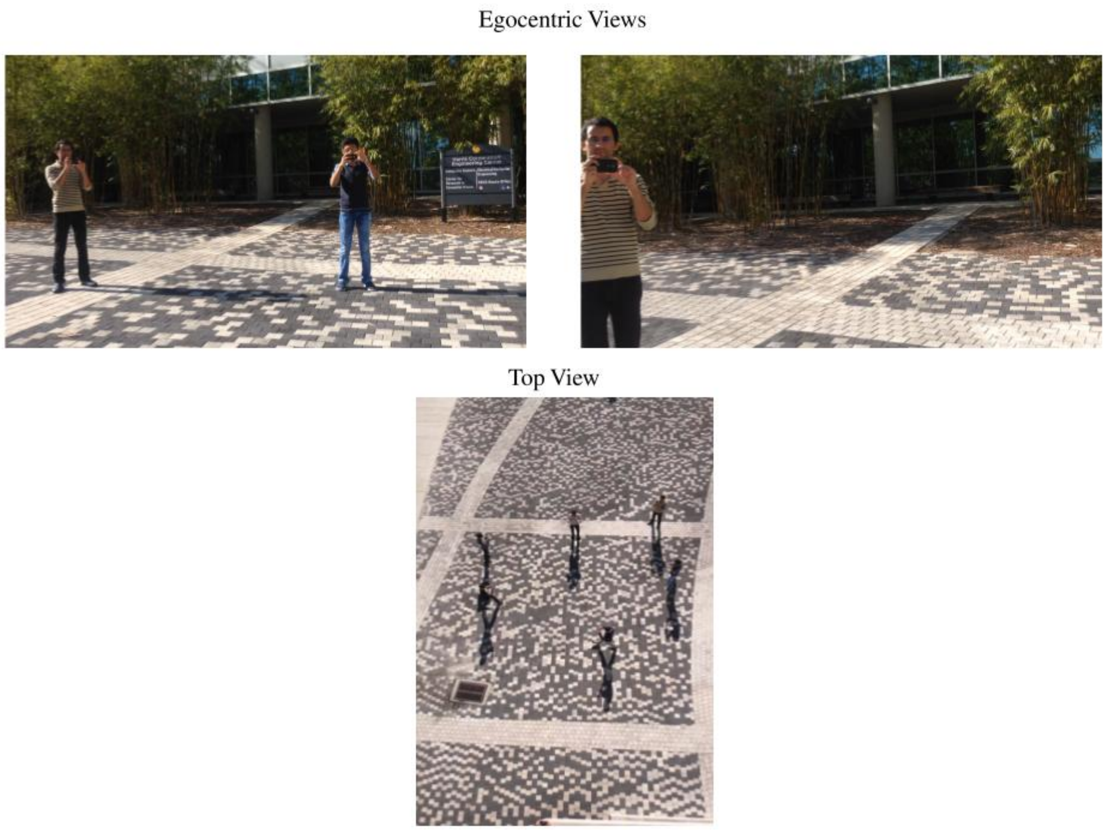


Figure 3.2: Ego and corresponding exo-view images taken from Ego2Top [6]

.

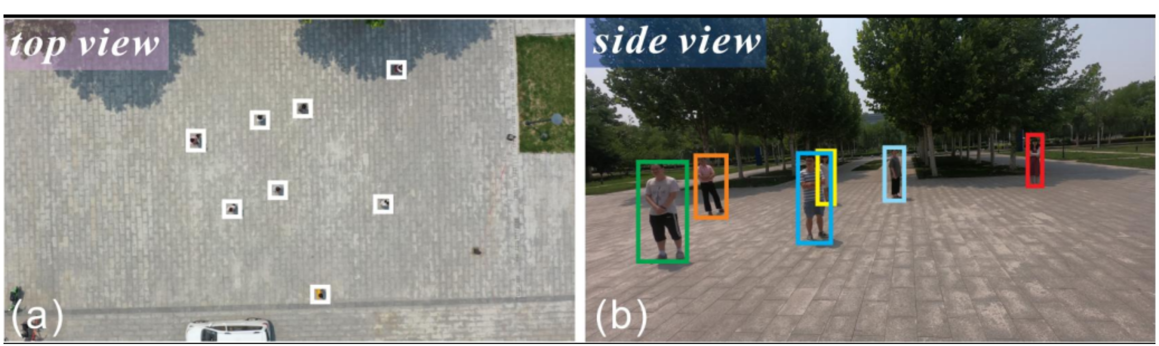


Figure 3.3: Top and side view-images taken from DMHA dataset released in [48]. Image taken from [48].

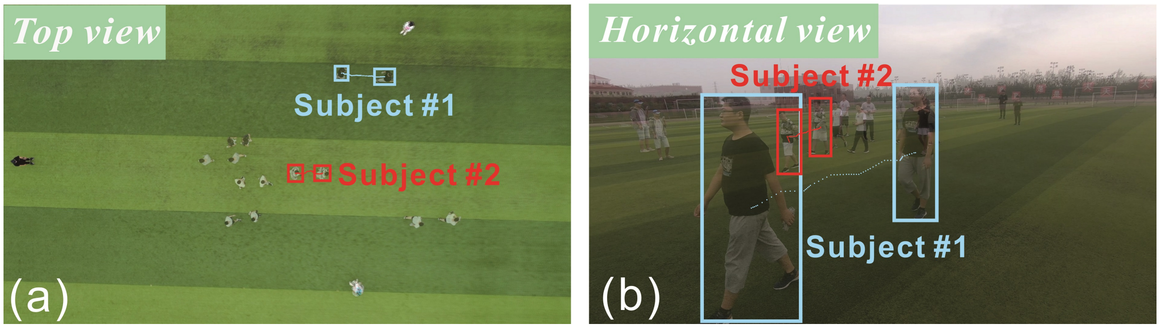


Figure 3.4: Example ego-exo frames taken from the CVMHT dataset released in [47]. Image taken from [47].


Figure 3.5: Pairs from simultaneously recorded Ego-Top and Ego-Side dataset. Image taken from [34].

Figure 3.6: Frames from the Demo2Vec paper

# Chapter 4

# Methodology

In this section, we describe our approach for recognizing egocentric actions using prior exocentric knowledge. We leverage the large-scale information encoded in vision-language models [107] and large language models [155, 135] for solving the task of egocentric action recognition. Our approach is illustrated in the figure 4.1.

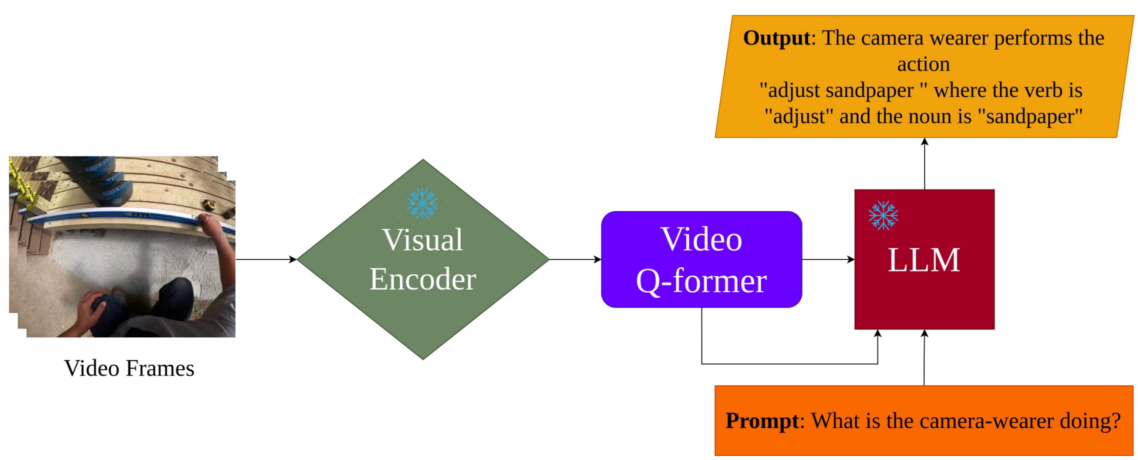


Figure 4.1: This presents a figure of our approach. Firstly, a frozen visual encoder extracts features. Then, a trainable video Q-former extracts $k$ queries and project to LLM's embedding space. Finally, the visual prompts and concatenated with the textual prompts and passed to the LLM.

### 4.0.1 Problem Definition

We describe the problem statement of egocentric action recognition. We follow the same approach as given in [24]. Formally, given a video clip $A$, we aim to classify to the action class, which is a tuple $C_a = \{(c_v, c_n)\}$, where $c_v \in C_v$ is the possible set of verbs and $c_n \in C_n$ is the possible set of nouns. For an accurate classification, we want to correctly predict both the verb and noun. Top-1 accuracy is used as a metric. A good survey of previous approaches can be found in Nunez *et. al* [98]. While previous approaches have achieved good verb-level accuracy, a significant gap exists in the noun-level and action-level accuracies.

### 4.0.2 Network architecture

Our architecture is based on the BLIP-2 architecture [74] and [153]. BLIP-2 proposes a novel and efficient training strategy, bootstrapping from a pretrained CLIP-based image encoder [107] and LLM. We use a hand-crafted prompt for the task of action recognition. It is a computationally efficient approach focusing on only training a lightweight Querying Transformer. It is pre-trained on a large-scale Internet-image dataset of 129M images, mostly composed of exocentric images. We hypothesize that exo-pre-trained VLM encodes useful signals that can be transferrable to egocentric images. Similar approaches are covered in **??**. Our model consists of three major components: 1) a pretrained frozen visual encoder, 2) Lightweight Trainable Video Query-former and 3) Frozen LLM. We describe the components below in more detail.

1) **Visual encoder.** This is a ViT-L/14 CLIP-based image-encoder, a dark-green

block in the 4.1. We sample $T$ frames from a video (usually chosen to be 8/16). They are preprocessed to a shape of $(224 \times 224)$ before passing into the image encoder. The last layer of the ViT is used, having shape $(P * T \times D)$, where $P$ is the number of patches and $D$ is the dimension of the image-encoder. The output of this layer is $V \in \mathbb{R}^{B \times K \times D}$, where $K$ is the product of the number of patches across all the frames and $B$ is the batch-size.

2) **Video Q-former.** It is an attention-based transformer [137] that generates a visual representation via self-attention between the shared layers. For a detailed overview, check the BLIP-2 paper by Li. *et. al* [74]. A fixed number of query tokens, $t$ are learned. $V$ is input to this layer and we get an output $Q \in \mathbb{R}^{B \times t \times D}$. We set $t$ as 32 and the dimension $D$ as 768.

3) **LLM.** We use a pre-trained, frozen LLM, OPT-175B [155] from Meta. LLMs encode commonsense knowlede that we can utilize for the task of action recognition. The previous layer's input $Q$ is projected to the LLM's embedding space, $I \in \mathbb{R}^{B \times t \times D}$, where $D$ is the LLM's input-embedding dimension. The textual prompt "What is the camera wearer doing?" is also converted to LLM's embedding space. $J \in \mathbb{R}^{B \times N \times D}$, where $N$ is the number of input ids obtained from the input text. We concatenate $I$ and $J$ and pass it to the LLM.

The model is end-to-end trained using a cross-entropy loss with the next-token prediction objective for the LLM. However, only the parameters of Video Q-former are trained.

### 4.0.3 Experiments

We provide a comprehensive overview to the various experiments that we performed and compare with the existing state-of-the art methodologies. We analyze the results thoroughly and understand the gaps.

We conduct evaluations on the two largest egocentric action recognition datasets: Ego4D [43] and EPIC-Kitchens100[23]. The top-1 verb, noun and action accuracy is used for the comparison. Refer to Tables 4.1 and 4.2 for results.

Table 4.1: Comparison of our technique vs the existing state-of-the-art methods on Ego4D action recognition

| Method | Prior Exo knowledge | Verb Acc. | Noun Acc. | Action Acc. |
|---|---|---|---|---|
| MViT [37] | ✓ | 19.87 | 2.55 | 0.51 |
| SlowFast [39] | ✓ | 19.42 | 14.65 | 3.12 |
| StillFast [108] | ✓ | 19.12 | 19.41 | 4.06 |
| **OURS + full response** | ✓ | 24.56 | 37.62 | 10.88 |
| **OURS + action response** | ✓ | 24.74 | 37.1 | 11.52 |
| EgoVLPv2 [104] | × | 33.84 | 40.63 | 16.31 |
| EgoVLP [77] | × | 40.32 | 45.53 | 20.63 |

In the table 4.1, **OURS + full response** means that the LLM response is the entire sentence *The camera wearer is performing the action "wear shirt", where the verb is "wear" and "noun" is shirt* and the **OURS + action response** forces the LLM to predict *wear shirt*. We can see a slight improvement when we predict only the verb-noun pair rather than the entire sentence.

From the table 4.1, we can see that our technique outperforms other exo-to-ego transfer techniques. This is due to the large-scale information embedded in VLM and LLMs. However, we are still behind EgoVLP [77] and EgoVLPv2 [104], that are large-scale

Table 4.2: Comparison of our technique vs the existing state-of-the-art methods on EPIC-Kitchens100 action recognition

| Method | Verb Acc. | Noun Acc. | Action Acc. |
|---|---|---|---|
| **OURS + full response** | 24.86 | 34.68 | 12.8 |
| TAdaFormer-L/14 [58] | 71.7 | 64.1 | 51.8 |
| M&M [140] | 72 | 66.3 | 53.6 |
| Avion [158] | 73 | 65.4 | 54.4 |

egocentric pretrained models. We hypothesize that large-scale pretraining on egocentric data learns better egocentric cues for classification. Similarly, from table 4.2, we observe that Avion, a large-scale video pretrained model outperforms other approaches by a huge margin.

### 4.0.4 Analysis

We analyze the limitations behind our technique in this section. Some of the possible drawbacks are:

**a) noun recognition.** A huge drawback behind noun-based recognition is the difficulty to accurately predict nouns in the egocentric view. Since our model is trained on the objects captured from the third person view, detecting the same from the egocentric view is challenging. Additionally, challenges of occlusion and cluttered scene makes it difficult to accurately detect the objects. From (c) Figure 4.2, we can see that the ground-truth object **multimeter** is composed of many different parts and the model is unable to accurately detect it.

**b) verb recognition.** The (a) Figure 4.2 shows that the model struggles to identify the ground-truth verb *cut* and is confused by the background noise.

**c) action recognition.** Predicting both the verb and noun exactly in the case of egocentric action recognition is very challenging. Due to the presence of the motorcycle in (b) of Figure 4.2, the model misses the hand-object interaction with the bucket.


Figure 4.2: Analyzing the qualitative results of our method on Ego4D.

# Chapter 5

# Conclusion and Future Works

This thesis is an attempt to transfer exocentric knowledge for egocentric tasks. Large-scale exocentric pretrained VLMs contain relevant cues that can be transferrable to downstream egocentric tasks. In our work, we propose a computationally efficient model for the task of egocentric action recognition. While, we are able to outperform previous exo-to-ego transfer techniques for egocentric action recognition, we lag behind pure ego-to-ego methods. In the future work, we will focus on utilizing the paired ego-exo data present in the new datasets [44, 57, 76] and focus on learning hand-object centric cues during training.

# Bibliography


[1] Kinetics-700 classification results. https://paperswithcode.com/sota/action-classification-on-kinetics-700. Accessed: 2024-07-02.

[2] Adnan Abdullah, Ruo Chen, Ioannis Rekleitis, and Md Jahidul Islam. Ego-to-exo: Interfacing third person visuals from egocentric views in real-time for improved rov teleoperation. *arXiv preprint arXiv:2407.00848*, 2024.

[3] Khush Agrawal and Rohit Lal. Person following mobile robot using multiplexed detection and tracking. In *Advances in Mechanical Engineering: Select Proceedings of ICAME 2020*, pages 815–822. Springer Singapore Singapore, 2020.

[4] Khush Agrawal, Rohit Lal, Himanshu Patil, Surender Kannaiyan, and Deep Gupta. Deepsct: Deep learning based self correcting object tracking mechanism. In *2021 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2021.

[5] Alexandre Alahi, Michel Bierlaire, and Murat Kunt. Object detection and matching with mobile cameras collaborating with fixed cameras. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications-M2SFA2 2008*, 2008.

[6] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 253–268. Springer, 2016.

[7] Shervin Ardeshir and Ali Borji. Egocentric meets top-view. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1353–1366, 2018.

[8] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171:61–68, 2018.

[9] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint identification and temporal alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] Shervin Ardeshir and Ali Borji. Integrating egocentric videos in top-view surveillance videos: Joint widentification and temporal alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 285–300, 2018.

[11] Shervin Ardeshir, Krishna Regmi, and Ali Borji. Egotransfer: Transferring motion across egocentric and exocentric domains using deep neural networks. *arXiv preprint arXiv:1612.05836*, 2016.

[12] Shervin Ardeshir, Sandesh Sharma, and Ali Broji. Egoreid: Cross-view self-identification and human re-identification in egocentric and surveillance videos. *arXiv preprint arXiv:1612.08153*, 2016.

[13] Shervin Ardeshir Behrostaghi. Relating first-person and third-person vision. 2018.

[14] Paola Ardón, Èric Pairet, Katrin S Lohan, Subramanian Ramamoorthy, and Ronald Petrick. Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*, 2020.

[15] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, 2023.

[16] Md Mushfiqur Azam and Kevin Desai. A survey on 3d egocentric human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1654, 2024.

[17] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[18] Emrah Basaran, Yonatan Tariku Tesfaye, and Mubarak Shah. Egoreid dataset: person re-identification in videos acquired by mobile devices with first-person point-of-view. *arXiv preprint arXiv:1812.09570*, 2018.

[19] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021.

[20] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset, 2022.

[21] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2023.

[22] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1060–1068, 2021.

[23] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55, 2022.

[24] 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 *ECCV*, 2018.

[25] Chris Dede. Introduction to virtual reality in education. *Themes in Science and Technology Education*, 2:7–9, 2009.

[26] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2016.

[27] Ameya Dhamanaskar, Mariella Dimiccoli, Enric Corona, Albert Pumarola, and Francesc Moreno-Noguer. Enhancing egocentric 3d pose estimation with third person views. *Pattern Recognition*, 138:109358, 2023.

[28] Sourish Gunesh Dhekane and Thomas Ploetz. Transfer learning in human activity recognition: A survey. *arXiv preprint arXiv:2401.10185*, 2024.

[29] Fangqiang Ding, Yunzhou Zhu, Xiangyu Wen, and Chris Xiaoxuan Lu. Thermohands: A benchmark for 3d hand pose estimation from egocentric thermal image. *arXiv preprint arXiv:2403.09871*, 2024.

[30] Stuart Duncan, Holger Regenbrecht, and Tobias Langlotz. Fusing exocentric and egocentric real-time reconstructions for embodied immersive experiences. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–11. IEEE, 2023.

[31] Matt Dunleavy and Chris Dede. Augmented reality teaching and learning. *Handbook of research on educational communications and technology*, pages 735–745, 2014.

[32] Arindam Dutta, Rohit Lal, Yash Garg, Calvin-Khang Ta, Dripta S Raychaudhuri, Hannah Dela Cruz, and Amit K Roy-Chowdhury. Posture: Pose guided unsupervised domain adaptation for human body part segmentation. *arXiv preprint arXiv:2407.03549*, 2024.

[33] Arindam Dutta, Rohit Lal, Dripta S. Raychaudhuri, Calvin-Khang Ta, and Amit K. Roy-Chowdhury. Poise: Pose guided human silhouette extraction under occlusions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6153–6163, January 2024.

[34] Mohamed Elfeki, Krishna Regmi, Shervin Ardeshir, and Ali Borji. From third person to first person: Dataset and baselines for synthesis and retrieval. *arXiv preprint arXiv:1812.00104*, 2018.

[35] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[36] Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J Crandall, and Michael S Ryoo. Identifying first-person camera wearers in third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5125–5133, 2017.

[37] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[38] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018.

[39] 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.

[40] John H. Flavell, Eleanor R. Flavell, Frances L. Green, and Sharon A. Wilcox. The development of three spatial perspective-taking rules. *Child Development*, 1981.

[41] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[42] Raphaël Grasset, Philip Lamb, and Mark Billinghurst. Evaluation of mixed-space collaboration. In *Fourth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'05)*, pages 90–99. IEEE, 2005.

[43] 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.

[44] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[45] Ruize Han, Wei Feng, Feifan Wang, Zekun Qian, Haomin Yan, and Song Wang. Benchmarking the complementary-view multi-human association and tracking. *International Journal of Computer Vision*, 132(1):118–136, 2024.

[46] Ruize Han, Wei Feng, Yujun Zhang, Jiewen Zhao, and Song Wang. Multiple human association and tracking from egocentric and complementary top views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5225–5242, 2021.

[47] Ruize Han, Wei Feng, Jiewen Zhao, Zicheng Niu, Yujun Zhang, Liang Wan, and Song Wang. Complementary-view multiple human tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10917–10924, 2020.

[48] Ruize Han, Yiyang Gan, Jiacheng Li, Feifan Wang, Wei Feng, and Song Wang. Connecting the complementary-view videos: joint camera identification and subject association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2416–2425, 2022.

[49] Ruize Han, Yiyang Gan, Likai Wang, Nan Li, Wei Feng, and Song Wang. Relating view directions of complementary-view mobile cameras via the human shadow. *International Journal of Computer Vision*, 131(5):1106–1121, 2023.

[50] Ruize Han, Yujun Zhang, Wei Feng, Chenxing Gong, Xiaoyu Zhang, Jiewen Zhao, Liang Wan, and Song Wang. Multiple human association between top and horizontal views by matching subjects' spatial distributions. *arXiv preprint arXiv:1907.11458*, 2019.

[51] Ruize Han, Jiewen Zhao, Wei Feng, Yiyang Gan, Liang Wan, and Song Wang. Complementary-view co-interest person detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2746–2754, 2020.

[52] Shengyu Hao, Peiyuan Liu, Yibing Zhan, Kaixun Jin, Zuozhu Liu, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang. Divotrack: A novel dataset and baseline method for cross-view multi-object tracking in diverse open scenes. *International Journal of Computer Vision*, 132(4):1075–1090, 2024.

[53] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018.

[54] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[55] Kyle Hsu, Moo Jin Kim, Rafael Rafailov, Jiajun Wu, and Chelsea Finn. Vision-based manipulators need to also see from their hands. *arXiv preprint arXiv:2203.12677*, 2022.

[56] Yi Huang, Xiaoshan Yang, Junyun Gao, and Changsheng Xu. Holographic feature learning of egocentric-exocentric videos for multi-domain action recognition. *IEEE Transactions on Multimedia*, 24:2273–2286, 2021.

[57] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024.

[58] Ziyuan Huang, Shiwei Zhang, Liang Pan, Zhiwu Qing, Yingya Zhang, Ziwei Liu, and Marcelo H Ang Jr. Temporally-adaptive models for efficient video understanding. *arXiv preprint arXiv:2308.05787*, 2023.

[59] Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022.

[60] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020.

[61] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[62] Wenqi Jia, Miao Liu, Hao Jiang, Ishwarya Ananthabhotla, James M Rehg, Vamsi Krishna Ithapu, and Ruohan Gao. The audio-visual conversational graph: From an egocentric-exocentric perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26396–26405, 2024.

[63] XIN Jianjia, WANG Lichun, and YIN Baocai. Survey of research on affordance understanding based on computer vision. *Journal of Beijing University of Technology*.

[64] Soo-Han Kang and Ji-Hyeong Han. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *International Journal of Social Robotics*, 15(4):631–641, 2023.

[65] Misha Karim, Shah Khalid, Aliya Aleryani, Jawad Khan, Irfan Ullah, and Zafar Ali. Human action recognition systems: A review of the trends and state-of-the-art. *IEEE Access*, 2024.

[66] William G Kennedy, Magdalena D Bugajska, Matthew Marge, William Adams, Benjamin R Fransen, Dennis Perzanowski, Alan C Schultz, and J Gregory Trafton. Spatial representation and reasoning for human-robot collaboration. In *AAAI*, volume 7, pages 1554–1559, 2007.

[67] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[68] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[69] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021.

[70] F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. In *Tech. report CMU-RI-TR-08-22, Robotics Institute, Carnegie Mellon University*, 2009.

[71] Rohit Lal, Yash Garg, Arindam Dutta, Calvin-Khang Ta, Dripta S Raychaudhuri, M Salman Asif, and Amit K Roy-Chowdhury. Temp3d: Temporally continuous 3d human pose estimation under occlusions. *arXiv preprint arXiv:2312.16221*, 2023.

[72] Hannah Larson and Leia Stirling. Examination of human spatial reasoning capability and simulated autonomous rendezvous and docking monitoring performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 67, pages 465–470. SAGE Publications Sage CA: Los Angeles, CA, 2023.

[73] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023.

[74] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[75] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.

[76] Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding, 2024.

[77] 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.

[78] Zheyuan Lin, Shanshan Ji, Wen Wang, Mengjie Qin, Rong Yang, Minhong Wan, Jason Gu, Te Li, and Chunlong Zhang. A joint tracking system: Robot is online to access surveillance views. In *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1–6. IEEE, 2023.

[79] Gaowen Liu, Hugo Latapie, Ozkan Kilic, and Adam Lawrence. Parallel generative adversarial network for third-person to first-person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1917–1923, 2022.

[80] Gaowen Liu, Hao Tang, Hugo Latapie, and Yan Yan. Exocentric to egocentric image generation via parallel generative adversarial network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1843–1847. IEEE, 2020.

[81] Gaowen Liu, Hao Tang, Hugo M Latapie, Jason J Corso, and Yan Yan. Cross-view exocentric to egocentric video synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 974–982, 2021.

[82] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yijun Chen, Yao Guo, and Guang-Zhong Yang. Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. *IEEE Transactions on Multimedia*, 25:8880–8891, 2023.

[83] Yichao Lu, Cheng Chang, Himanshu Rai, Guangwei Yu, and Maksims Volkovs. Multi-view scene graph generation in videos. In *International Challenge on Activity Recognition (ActivityNet) CVPR 2021 Workshop*, volume 3, page 2, 2021.

[84] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[85] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *International Journal of Computer Vision*, pages 1–25, 2023.

[86] Hongchen Luo, Kai Zhu, Wei Zhai, and Yang Cao. Intention-driven ego-to-exo video generation. *arXiv preprint arXiv:2403.09194*, 2024.

[87] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. *arXiv preprint arXiv:2403.06351*, 2024.

[88] Nikoleta Manakitsa, George S Maraslidis, Lazaros Moysis, and George F Fragulis. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. *Technologies*, 12(2):15, 2024.

[89] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[90] Paul Milgram, Herman Colquhoun, et al. A taxonomy of real and virtual world display integration. *Mixed reality: Merging real and virtual worlds*, 1(1999):1–26, 1999.

[91] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[92] Xianliang Mu, Lifen Tan, Yu Tian, and Chunhui Wang. The effect of multiple perspectives information on the characteristics of human's spatial cognition in the human-human interaction of spatial cognition tasks. In *Engineering Psychology and Cognitive Ergonomics: 13th International Conference, EPCE 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings 13*, pages 69–78. Springer, 2016.

[93] Neeti Narayan, Nishant Sankaran, Srirangaraj Setlur, and Venu Govindaraju. Learning deep features for online person tracking using non-overlapping cameras: A survey. *Image and Vision Computing*, 89:222–235, 2019.

[94] Nora Newcombe. The development of spatial perspective taking. *Advances in child development and behavior*, 1989.

[95] Jing Ng, David Arness, Ashlee Gronowski, Zhonglin Qu, Chng Wei Lau, Daniel Catchpoole, and Quang Vinh Nguyen. Exocentric and egocentric views for biomedical data analytics in virtual environments—a usability study. *Journal of Imaging*, 10(1), 2024.

[96] Bahareh Nikpour, Dimitrios Sinodinos, and Narges Armanfard. Deep reinforcement learning in human activity recognition: A survey and outlook. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[97] Mai Nishimura, Shohei Nobuhara, and Ko Nishino. View birdification in the crowd: Ground-plane localization from perceived movements. *International Journal of Computer Vision*, 131(8):2015–2031, 2023.

[98] Adrián Núñez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. *Neurocomputing*, 472:175–197, 2022.

[99] 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.

[100] Takehiko Ohkawa, Takuma Yagi, Taichi Nishimura, Ryosuke Furuta, Atsushi Hashimoto, Yoshitaka Ushiku, and Yoichi Sato. Exo2egodvc: Dense video captioning of egocentric procedural activities using web instructional videos. *arXiv preprint arXiv:2311.16444*, 2023.

[101] Gao Peng, Yong-Lu Li, Hao Zhu, Jiajun Tang, Jin Xia, and Cewu Lu. Vvs: Action recognition with virtual view synthesis. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 384–388. IEEE, 2021.

[102] Joshua M Peschel. Towards physical object manipulation by small unmanned aerial systems. In *2012 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–6. IEEE, 2012.

[103] 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*, pages 1–57, 2024.

[104] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023.

[105] Zekun Qian, Ruize Han, Wei Feng, and Song Wang. From a bird's eye view to see: Joint camera and subject registration without the camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 863–873, 2024.

[106] Camillo Quattrocchi, Antonino Furnari, Daniele Di Mauro, Mario Valerio Giuffrida, and Giovanni Maria Farinella. Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs. *arXiv preprint arXiv:2312.02638*, 2023.

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

[108] 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 3636–3645, 2023.

[109] Arushi Rai, Kyle Buettner, and Adriana Kovashka. Strategies to leverage foundational model knowledge in object affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1714–1723, 2024.

[110] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, and J.C. Niebles. Home action genome: Contrastive compositional action understanding. In *CVPR*, 2021.

[111] Karinne Ramirez-Amaro, Humera Noor Minhas, Michael Zehetleitner, Michael Beetz, and Gordon Cheng. Added value of gaze-exploiting semantic representation to allow robots inferring human behaviors. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–30, 2017.

[112] Troels Ammitsbøl Rasmussen and Weidong Huang. Scenecam: Improving multi-camera remote collaboration using augmented reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 28–33. IEEE, 2019.

[113] Troels Ammitsbøl Rasmussen and Weidong Huang. Scenecam: Improving multi-camera remote collaboration using augmented reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 28–33, 2019.

[114] Dripta S. Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, and Amit K. Roy-Chowdhury. Prior-guided source-free domain adaptation for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14996–15006, October 2023.

[115] Carmine Tommaso Recchiuto, Antonio Sgorbissa, and Renato Zaccaria. Visual feedback with multiple cameras in a uavs human–swarm interface. *Robotics and Autonomous Systems*, 80:43–54, 2016.

[116] C.T. Recchiuto, A. Sgorbissa, and R. Zaccaria. Visual feedback with multiple cameras in a uavs human–swarm interface. *Robotics and Autonomous Systems*, 80:43–54, 2016.

[117] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018.

[118] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[119] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27(1):169–192, 2004.

[120] Bernardo Rocha, Plinio Moreno, and Alexandre Bernardino. Cross-view generalisation in action recognition: Feature design for transitioning from exocentric to egocentric views. In *Iberian Robotics conference*, pages 155–166. Springer, 2023.

[121] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252, 2021.

[122] 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.

[123] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[124] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018.

[125] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[126] Leonardo Pavanatto Soares, Regis Kopper, and Márcio Sarroglia Pinho. Ego-exo: A cooperative manipulation technique with automatic viewpoint control. In *2018 20th Symposium on Virtual and Augmented Reality (SVR)*, pages 82–88. IEEE, 2018.

[127] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

[128] Bilge Soran, Ali Farhadi, and Linda Shapiro. Action recognition in the presence of one egocentric and multiple static cameras. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, pages 178–193. Springer, 2015.

[129] Josua Spisak, Matthias Kerzel, and Stefan Wermter. Diffusing in someone else's shoes: Robotic perspective taking with diffusion. *arXiv preprint arXiv:2404.07735*, 2024.

[130] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024.

[131] Hao Tang, Philip HS Torr, and Nicu Sebe. Multi-channel attention selection gans for guided image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6055–6071, 2022.

[132] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2417–2426, 2019.

[133] Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The coin dataset and performance evaluation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[134] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021.

[135] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[136] Thanh-Dat Truong and Khoa Luu. Cross-view action recognition understanding from exocentric to egocentric perspective. *arXiv preprint arXiv:2305.15699*, 2023.

[137] 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.

[138] Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, and Xi Peng. Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3307–3317, 2023.

[139] Yangming Wen, Krishna Kumar Singh, Markham Anderson, Wei-Pang Jan, and Yong Jae Lee. Seeing the unseen: Predicting the first-person camera wearer's location

and pose in third-person scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3446–3455, 2021.

[140] Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&m mix: A multimodal multiview transformer ensemble. *arXiv preprint arXiv:2206.09852*, 2022.

[141] Boshen Xu, Sipeng Zheng, and Qin Jin. Pov: Prompt-oriented view-agnostic learning for egocentric hand-object interaction in the multi-view world. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2807–2816, 2023.

[142] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024.

[143] Lingjing Xu, Yang Gao, Wenfeng Song, and Aimin Hao. Weakly supervised multimodal affordance grounding for egocentric images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6324–6332, 2024.

[144] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S Ryoo, and David J Crandall. Joint person segmentation and identification in synchronized first-and third-person videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–652, 2018.

[145] Zihui Sherry Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. *Advances in Neural Information Processing Systems*, 36:53688–53710, 2023.

[146] Takuma Yagi, Misaki Ohashi, Yifei Huang, Ryosuke Furuta, Shungo Adachi, Toutai Mitsuyama, and Yoichi Sato. Finebio: A fine-grained video dataset of biological experiments with hierarchical annotations. In *Proceedings of the Joint International 3rd Ego4D and 11th EPIC Workshop (in conjunction with CVPR 2023, extended abstract)*, 2023.

[147] Fan Yang, Wenrui Chen, Kailun Yang, Haoran Lin, DongSheng Luo, Conghui Tang, Zhiyong Li, and Yaonan Wang. Learning granularity-aware affordances from human-object interaction for tool-based functional grasping in dexterous robotics. *arXiv preprint arXiv:2407.00614*, 2024.

[148] Liang Yang, Hao Jiang, Zhouyuan Huo, and Jizhong Xiao. Visual-gps: ego-downward and ambient video based person location association. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[149] Sierra N Young, Ryan J Lanciloti, and Joshua M Peschel. The effects of interface views on performing aerial telemanipulation tasks using small uavs. *International Journal of Social Robotics*, 14(1):213–228, 2022.

[150] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1358–1366, 2019.

[151] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. First-and third-person video co-analysis by learning spatial-temporal joint attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6631–6646, 2020.

[152] Jeongmin Yu, Seungtak Noh, Youngkyoon Jang, Gabyong Park, and Woontack Woo. A hand-based collaboration framework in egocentric coexistence reality. In *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 545–548. IEEE, 2015.

[153] Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, and Joyce Chai. Efficient in-context learning in vision-language models for egocentric videos. *arXiv preprint arXiv:2311.17041*, 2023.

[154] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022.

[155] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[156] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013.

[157] Zhipeng Zhang, Zhimin Wei, Guolei Sun, Peng Wang, and Luc Van Gool. Self-explainable affordance learning with embodied caption. *arXiv preprint arXiv:2404.05603*, 2024.

[158] Yue Zhao and Philipp Krähenbühl. Training a large video model on a single machine in a day. *arXiv preprint arXiv:2309.16669*, 2023.

[159] Ziwei Zhao, Yuchen Wang, and Chuhua Wang. Fusing personal and environmental cues for identification and segmentation of first-person camera wearers in third-person views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16477–16487, 2024.

[160] L. Zhou, N. Louis, and J. Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. In *BMVC*, 2018.

[161] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019.