Audio-Visual State-Aware Representation Learning from Interaction-Rich, Egocentric Videos

Himangi Mittal

CMU-RI-TR-23-09
April 2023

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Abhinav Gupta, chair
David Held
Shubham Tulsiani
Yufei Ye

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics.

To my loving family, friends, and mentors.

Abstract

We propose a self-supervised algorithm to learn representations from egocentric video data using multiple modalities of video and audio. In robotics and augmented reality, the input to the agent is a long stream of video from the first-person or egocentric point of view. Towards this end, recently there have been significant efforts to capture humans from their first-person/egocentric view interacting with their own environments as they go about their daily activities. As a result, several large-scale egocentric, interaction-rich, multi-modal datasets have emerged. However, learning representations from such videos can be quite challenging.

First, given the uncurated nature of long, untrimmed, continuous videos, learning effective representations require focusing on moments in time when interactions take place. A real-world video consists of many non-activity segments which are not conducive to learning. Second, visual representations of daily activities should be sensitive to changes in the state of the object and the environment. In other words, the representations should be state-aware. However, current successful multi-modal learning frameworks encourage representations that are invariant to time and object states.

To address these challenges, we leverage audio signals to identify moments of likely interactions which are conducive to better learning. Motivated by the observation of a sharp audio signal associated with an interaction, we also propose a novel self-supervised objective that learns from audible state changes caused by interactions. We validate these contributions extensively on two large-scale egocentric datasets, EPIC-Kitchens-100 and Ego4D, and show improvements on several downstream tasks, including action recognition, long-term action anticipation, object state change classification, and point-of-no-return temporal localization.

Acknowledgments

I would first like to thank my advisor, Prof. Abhinav Gupta for his guidance and support throughout my research and Master's. His passion, insights, and vision of robotics has deeply influenced my way to think about the research challenges in AI. A huge thanks to Prof. Pedro Morgado for constantly guiding me, supporting me, understanding me with patience, for his time, and fun chats. I would especially like to thank Abhinav and Pedro for introducing me to the wonderful research domain of multi-modal machine learning. I would also like to thank Unnat Jain and Senthil Purushwalkam for their fruitful feedback and for project discussions, especially near the submission deadline. Lastly, I would like to express my immense gratitude to all of them for guiding me towards my first NeurIPS publication.

I would also like to thank Prof. David Held, Prof. Shubham Tulsiani, and Yufei Ye for being my committee members. I also feel very fortunate to work with a great group of people in my lab - Yufei Ye, Shikhar Bahl, Homanga Bharadwaj, Victoria Dean, Raunaq Bhirangi, Sam Powers, and Sudeep Dasari. I would also like to thank Jared Mejia, Mohan Kumar, Ryan Aponte, and Gaoyue Zhou for casual chats and fun in the lab.

Lastly, I would like to thank my loving parents and my dear brother for their constant love and support. I would especially like to thank them for understanding me and respecting my decisions. Their love gives me the strength to work hard and to stay determined to my goals.

Funding

We would like to thank DARPA MCS, ONR Young Investigator and DARPA SAIL-ON for the funding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Self-supervised learning has witnessed tremendous progress with the help of pretext tasks [18, 27, 54, 57] or contrastive learning based methods [12, 30, 34, 55]. However, most of the methods are bottlenecked by the lack of rich, real-world data and they learn from static images which lack temporal information and restrict the ability to learn object deformations and state changes over time. This issue can be addressed by leveraging videos that provide temporal information and learning rich representations from them in a self-supervised manner.

Learning representations from videos is however quite challenging. We first need to choose the right objective function for self-supervised representation learning. Approaches such as [61, 74] learn representations that are invariant to object deformations and viewpoints. However, many downstream tasks such as action recognition require representations that are sensitive to these deformations in order to uniquely identify an action. Another alternative has been to use the multi-modal data [3, 50, 64] and learn representations via audio. However, most of these approaches seek to align audio and visual features in a common space which leads to invariant representations as well. The second challenge is that current video-based SSL approaches train on the curated nature of video datasets, such as Kinetics [10]. These approaches are designed to leverage carefully selected clips, displaying a single action or object interaction. This is in contrast to the predominantly *untrimmed* real-world data which consists of multiple daily activities and interactions over time. Here, unlike action centric datasets, the most ‘interesting‘ or ‘interaction-rich‘ clips have NOT been carefully selected by human annotators. A long, untrimmed video can consist of long non-activity segments. Thus, learning from untrimmed videos poses a major challenge, as a significant portion of the data does not focus on the concepts we want to learn.

At the same time, we see our world from an egocentric, first-person view. Similarly, for robotic agents, the input consists of a long stream of video from a first-person, egocentric point of view. Motivated by this, there have been efforts towards learning from egocentric datasets and curating datasets [15, 16, 29] which represent the real-world.

Motivated by the above challenges and the characteristics of the real-world, in this work, we ask the question, 'Can we learn meaningful representations from interaction-rich, multi-modal streams of egocentric data?' Learning from continuous streams of data requires focusing on the right moments in time when the actual interactions are likely to occur. Consider, for example, the acts of opening a fridge, placing a pan on the stove, or cutting of vegetables. Actions like these create clear and consistent sound signatures due to the physical interaction between objects. These moments can be easily detected from audio alone and can be used to target training on interesting portions of the untrimmed videos. We show that even a simple spectrogram-based handcrafted detector is sufficient to identify interesting moments in time, and that representation learning benefits substantially from using them to sample training clips.

Prior work on audio-visual correspondence (AVC) [4, 17, 50] use the natural co-occurrence of sounds as a source of supervision and try to bring the modalities into a common feature space. However, since the AVC objective still favors invariance, the learned representations are not informative of the changes that happen over time (e.g., representations that can distinguish between closed and opened fridge, or vegetables before and after chopping them). To better capture state changes, we introduce a novel audio-visual self-supervised objective, in which audio representations at key moments in time are required to be informative of the *change* in the corresponding visual representations over time. The intuition behind this objective is that transitions between object states are often marked by characteristic sounds. Thus, models optimized under this objective would associate the distinct sounds not only with the objects themselves (as accomplished with AVC), but also with the transition between two different states of the object.

To this end, we introduce RepLAI – **Rep**resentation **L**earning from **A**udible **I**nteractions, a self-supervised algorithm for representation learning from videos of audible interactions. RepLAI uses the audio signals in two unique ways: (1) to identify moments in time that are conducive to better self-supervised learning and (2) to learn representations that focus on the visual state changes caused by audible interactions. We validate these contributions extensively on two egocentric datasets, EPIC-Kitchens-100 [16] and the recently released Ego4D [29], where we demonstrate the benefits of RepLAI for several downstream tasks, including action recognition, long term action anticipation, and object state change classification.

Chapter 2

Related Works

2.1 Self-supervised learning

Self-supervised learning methods learn representations from an unlabeled dataset. These methods can be divided into two categories: learning from pretext tasks and learning from contrastive learning based objective functions. Multiple pretext tasks in self-supervised learning have been explored such as solving jigsaw puzzle [54], patch location prediction [18], inpainting [57], and image rotation [27] prediction. The second category of contrastive learning learns representations with the help of data augmentation and creates positive and negative pairs for a data sample. The positive pair is brought closer via cosine similarity and negative pairs are brought far apart. This is explored as instance discrimination [9, 12, 30, 34, 55]. These methods have shown rapid progress in self-supervised learning for images. While these approaches explore spatial information of images, our method leverages the temporal information of videos.

2.2 Video representation learning

For self-supervised representation learning in videos, spatio-temporal pretext tasks are designed such as temporal order prediction [39, 47, 76, 77], predicting motion and appearance statistics [72], pace prediction [73], temporal cycle consistency [20, 75], and video colorization [71]. The second category in self-supervised learning of contrastive learning has also been widely explored for videos [24, 31, 32, 36, 62, 64, 79] with impressive results on action recognition tasks. These methods however learn representations that are invariant to spatio-temporal augmentations, such as temporal jittering, and thus are incapable of representing object state changes. Closer to the objective of our method, we include relevant literature on audio-visual representation learning from videos, where the audio stream is additionally utilized.

2.3 Audio-visual representation learning

Leveraging other modalities to provide a supervisory signal has also been explored in the context of the audio modality with the help of audio-visual correspondence (AVC) [4, 5]. Pretext tasks for audio-visual representation learning include temporal synchronization [40, 56] between audio and video, audio classification [3, 6, 13], spatial alignment prediction between audio and 360-degree videos [48], optimal combination of self-supervised tasks [59]. The above tasks have been shown beneficial for learning effective multi-modal video representations. Contrastive learning has also been explored for both audio and video modality [49, 50, 58] as a cross-modal instance discrimination task. We explore the audio-visual representation for real-world fine-grained video understanding.

2.4 Fine-grained video understanding

Real-world videos are often long and untrimmed in nature and have multiple actions in a single video. Along this line, fine-grained analysis has been studied for videos in the form of a query-response temporal attention mechanism [80], bi-directional RNNs [65], and semi-supervised learning problem [19]. While these works are unimodal in nature and only utilize the visual modality, other works have also explored multi-modal fine-grained video understanding as a transformer-based model [38], by exploiting the correspondence between modalities [51], or by exploring how to best combine multiple modalities - audio, visual, and language [2]. In our work, we show fine-grained video understanding in a self-supervised manner by using the video and audio modalities.

2.5 Egocentric datasets

In robotics and augmented reality, the input to the agent is a long stream of video from the first-person or egocentric point of view. Motivated by this, efforts have been made toward curating egocentric datasets. These egocentric datasets offer new opportunities to learn from a first-person point of view, where the world is seen through the eyes of an agent. Many egocentric datasets have been developed such as Epic-kitchens [15, 16] which consist of daily activities performed in a kitchen environment, Activities of Daily Living [60], UT Ego [42, 67], the Disney Dataset [22], and the large-scale Ego4D dataset [29] which consists of daily life activities in multiple scenarios such as household, outdoor spaces, workplace, etc. Multiple challenges and downstream tasks have been explored for egocentric datasets like action recognition [37, 38, 43], action localization [63], action anticipation [1, 25, 28, 44, 66], human-object interactions [7, 14, 52], parsing social interactions [53], audio-visual navigation [11], and domain adaptation [51]. In our work, we evaluate the representations learned by our self-supervised approach on the EPIC-Kitchens-100 and Ego4D datasets over multiple downstream tasks.

2.6 Learning from interactions

Over the last few years, human-object interactions have been widely explored in the form of reconstructing the hand in the wild [8], from a single RGB image [78], reconstructing both hand and object from RGB videos [33], image generation [35], hand pose estimation [45], and two hand interactions [41]. Studying these interactions has been helpful for improving grasping [46]. While these works explore how to interact with the objects, in our work, we study the object state changes that arise from an interaction.

Chapter 3

Audio-visual State-Aware Representation Learning from Interactions

3.1 Objective

Our objective is to learn audio-visual representations from interaction-rich, ego-centric data, such that, it develops a state-aware understanding of the action/interaction being performed in a video clip. We achieve this in a self-supervised manner and make two key contributions in our work - **identifying moments of interaction (MoI)** and **learning from audible state changes**. In the following sections, we will provide an overview of our approach and then present our contributions in detail.

3.2 Overview

This section contains the setup of our algorithm for learning audio-visual representations. We use the framework of self-supervised learning to learn representations from audio-visual videos without annotations.

We begin with a dataset $D = \{(v_i, a_i)_{i=1}^{N}\}$ which consists of N long, untrimmed videos consisting of both video and audio modality. Given a sample $(v, a) \in D$, we

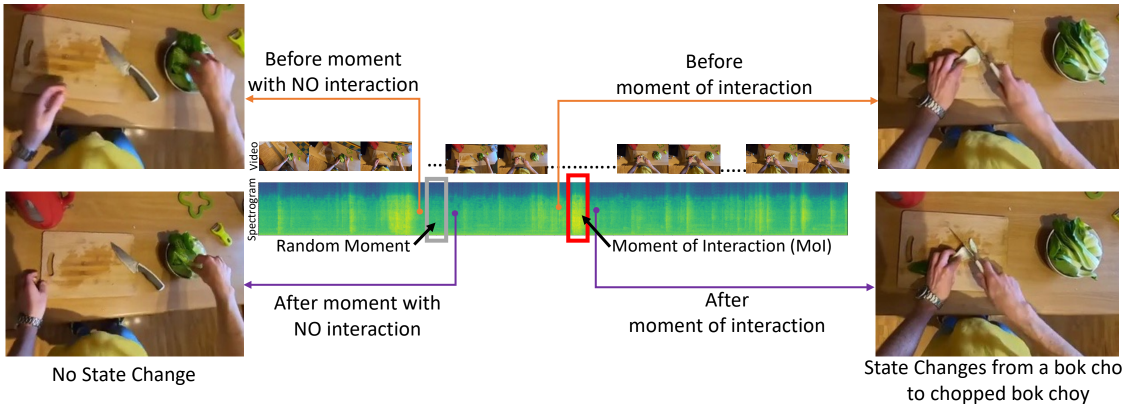

Figure 3.1: **Moment of interactions (MoI) detected in a long, untrimmed video.** When humans interact with the environment, it produces a unique audio pattern which is observed as *vertical edges* in the audio spectrogram. We define these edges as moments of interaction as shown in red box. Random moments in time are likely to contain NO interactions as shown by the gray box. Since no interactions occur, no changes are observed in the before and after states. We observe state changes around moments of interaction (MoI) which can be leveraged to learn state-aware representations.

identify the moment of interactions (MoI) using the audio signal of the video (Section 3.2.1). Then, we extract a short audio clip and short video clips around these identified moments of interaction. We use visual and audio encoders, f_V and f_A respectively, to encode these trimmed audio-video clips. Note that our objective is to learn the audio-visual encoders, f_V and f_A. Once we receive the audio-visual representations from the encoders, we optimize the algorithm with two audio-visual self-supervised losses:

1. Taking inspiration from the work [50], we use **Audio-Visual Correspondence Loss**, $\mathcal{L}_{\texttt{AVC}}$ (Section 3.2.3).
2. We propose a novel objective, **Audible State Change Loss** $\mathcal{L}_{\texttt{AStC}}$ (Section 3.2.2).

Before a detailed discussion of the moment of interaction and the objective functions used for model training, we provide some intuition behind our two key contributions as follows:

- **How do moments of interaction (MoI) help?** When we consider a real-

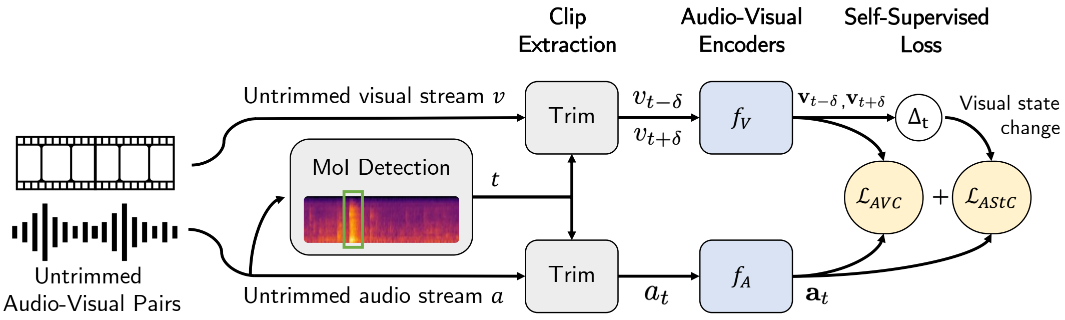

Figure 3.2: **Overview of RepLAI.** RepLAI seeks to learn audio and visual encoders (f_A and f_V) by (1) detecting and focusing the training on moments of interaction (MoI) present in untrimmed videos and (2) learning via two self-supervised objective functions – audio-visual correspondence (AVC) and audio identifiable state changes (AStC).

world, untrimmed video of daily activities, it often contains long periods without interactions, which aren't useful for training. Instead, we search for moments in time that are more likely to contain interactions, which we refer to as moments of interaction (MoI). This gives a more informative signal to the model to learn about the interactions with the environment. Note that our definition of an interaction is not just restricted to a human-object interaction but also covers the scenarios of human-environment interaction.

- **How does audible state change loss ($\mathcal{L}_{\texttt{AStC}}$) help?** Visual representations of interaction-rich data should be informative of the changes in the state of the environment and/or objects being interacted with. Moreover, these state changes are usually caused by physical interactions, which produce distinct sound signatures. We hypothesize that state-aware representations can be obtained by learning to associate audio with the change of visual representation during a moment of interaction.

3.2.1 Audio-driven identification of moments of interaction (MoI)

We hypothesize that audio signals can be particularly informative of moments of interaction. Considering a real-world scenario, when we perform day-to-day activities, we physically interact with the objects in our environment. These interactions usually produce a distinct audio pattern which is a short burst of energy that span all frequencies (Figure 3.1). Referring to Figure 3.1, we visualize the untrimmed visual and audio data of a person interacting with the environment. The audio modality is represented as a log mel spectrogram where the x-axis represents time and y-axis the audio frequency in log-scale. The moments of interaction are visible in the spectrogram in the form of *vertical edges* which can be easily detected to give us the timestamps of where an interaction occurred. Once detected, we take short audio-video clips around these moments of interaction and collect them into a dataset $\mathcal{D}_{\texttt{MoI}}$ which is used for training.

Now, we focus our discussion on *how we can locate the timestamp of such vertical edges*. Intuitively, we do this by finding robust local maxima in the total energy (summed over all frequencies) of the spectrogram. Concretely, let $M(t, \omega)$ be the value of the log mel spectrogram of an audio clip at time t and frequency ω. To remove the background noise, we compute the z-score normalization of the spectrogram for each frequency independently $\bar{M}(t, \omega) = \frac{s(t,\omega)-\mu_\omega}{\sigma_\omega+\epsilon}$, where ϵ is small constant for numerical stability. Here, μ_ω and σ_ω are the mean and standard deviation of $M(t, \omega)$ over time, respectively. [1]

Next, we define moments of interaction as the set of timestamps which are local maxima of $\sum_\omega \bar{s}(t, \omega)$ (or peaks for short). Since there can be multiple local maxima or weak local maxima due to the noisy nature of audio signals, we ignore peaks with small prominence (lower than 1)[2]. For further robustness, when multiple close peaks are found (less than 50ms apart), only the highest prominence peak is kept.

3.2.2 Audible State Change Loss ($\mathcal{L}_{\texttt{AStC}}$)

When we interact with the environment, the physical interactions often gives rise to two natural things:

- State Changes in the environment
- Distinct audio signals

[1] Specifically, $\mu_\omega = \mathbb{E}_t[M(t, \omega)]$, $\sigma_\omega^2 = \mathbb{E}_t[(M(t, \omega) - \mu_\omega)^2]$, and $\epsilon = 1e - 5$.

[2] The prominence of a peak is defined as the difference between the peak value and the minimum value in a small window around it.

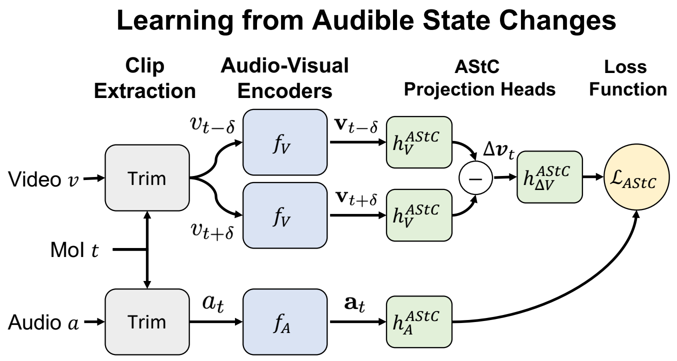

Figure 3.3: The proposed `AStC` formulation (3.2.2)

We leverage this natural co-occurrence and propose a self-supervised objective that seeks to *associate the audio with changes in the visual state* during a moment of interaction. The proposed task is optimized by minimizing a loss with two negative log-likelihood terms to:

- *increase* the probability of associating the audio with the visual state change in the *forward* (i.e. correct) direction
- *decrease* the probability of associating the audio with the visual state change in the *backward* (i.e. incorrect) direction

For example, consider the interaction of 'cutting a vegetable' as shown in the Figure 3.1. To optimize for this task, the audio of the action cut should be

- similar to the visual transition of *full vegetable* → *cut vegetable*
- dissimilar to the (backwards) transition *cut vegetable* → *full vegetable*

This encourages the model to learn representations that are informative of object states as well as the transition direction of the object states, making them useful for a variety of egocentric tasks. Specifically, the audible state change (`AStC`) loss can be defined as:

$$\mathcal{L}_{\texttt{AStC}} = \mathbb{E}_{v_t,a_t\in\mathcal{D}_{\texttt{MoI}}}\left[-\log\left(p^{\text{frwd}}(v_t,a_t)\right) - \log\left(1-p^{\text{bkwd}}(v_t,a_t)\right)\right]. \tag{3.1}$$

The probabilities (p^{frwd}, p^{bkwd}) are computed from cross-modal similarities

$$p^{\text{frwd}}(v_t, a_t) = \sigma\left(\texttt{sim}\left(\Delta\mathbf{v}_t^{\text{frwd}}, \mathbf{a}_t\right)/\tau\right) \tag{3.2}$$

$$p^{\text{bkwd}}(v_t, a_t) = \sigma\left(\texttt{sim}\left(\Delta\mathbf{v}_t^{\text{bkwd}}, \mathbf{a}_t\right)/\tau\right) \tag{3.3}$$

where $\tau = 0.2$ is a temperature hyper-parameter, and σ denotes the sigmoid function. For readability, we absorb the notations for the audio projection MLP head $h_A^{\texttt{AStC}}$ and the state change projection MLP head $h_{\Delta V}^{\texttt{AStC}}$ within $\texttt{sim}(\cdot, \cdot)$, but their usage is clearly illustrated in 3.3.

We obtain the audio representations ($\mathbf{a}_t$) by encoding the trimmed audio clips a_t via the audio encoder f_A (shared across all objectives). As explained above, $\mathbf{a}_t$ is further projected via $h_A^{\texttt{AStC}}$ to a space where the similarity to visual state changes is enforced.

State change representations ($\Delta\mathbf{v}_t^{\text{frwd}}$, $\Delta\mathbf{v}_t^{\text{bkwd}}$) are computed by considering two non-overlapping visual clips for each moment of interaction t, at timestamps $t - \delta$ and $t + \delta$. The two clips, $v_{t-\delta}$ and $v_{t+\delta}$, are encoded via the visual encoder f_V (shared across all tasks) and a projection MLP head $h_V^{\texttt{AStC}}$ (specific to the AStC task). Specifically, we represent forward and backward state changes as

$$\Delta\mathbf{v}_t^{\text{frwd}} = h_V^{\texttt{AStC}} \circ f_V(v_{t+\delta}) - h_V^{\texttt{AStC}} \circ f_V(v_{t-\delta}), \tag{3.4}$$

$$\Delta\mathbf{v}_t^{\text{bkwd}} = h_V^{\texttt{AStC}} \circ f_V(v_{t-\delta}) - h_V^{\texttt{AStC}} \circ f_V(v_{t+\delta}). \tag{3.5}$$

3.2.3 Audio-Visual Correspondence Loss ($\mathcal{L}_{\texttt{AVC}}$)

Audio-visual correspondence (AVC) is a well-studied self-supervised method [4, 17, 50] for learning uni-modal audio and visual encoders. The key idea is to bring visual and audio clips into a common feature space such that the representations of audio-visual pairs are aligned.

For our audio-visual correspondence objective (AVC), we consider a dataset of audio-visual pairs (v_i, a_i) with representations $\mathbf{v}_i = f_V(v_i)$ and $\mathbf{a}_i = f_A(a_i)$. In our dataset, (v_i, a_i) are short video and audio clips extracted from sample i around one of the detected moments of interaction. Then, taking inspiration from the work [50, 68], audio-visual correspondence is established by minimizing a cross-modal InfoNCE loss

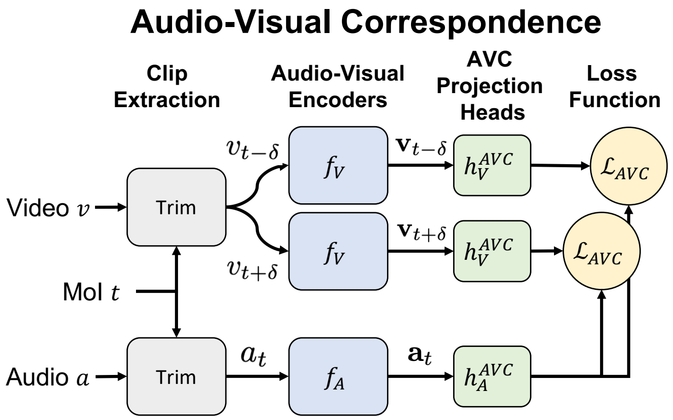

Figure 3.4: Audio-visual correspondence (3.2.3)

of the form as follows:

$$\mathcal{L}_{\texttt{AVC}} = \mathbb{E}_{v_i,a_i\sim D}\left[-\log\frac{e^{\texttt{sim}(\mathbf{v}_i,\mathbf{a}_i)/\tau}}{\sum_j e^{\texttt{sim}(\mathbf{v}_i,\mathbf{a}_j)/\tau}} - \log\frac{e^{\texttt{sim}(\mathbf{v}_i,\mathbf{a}_i)/\tau}}{\sum_j e^{\texttt{sim}(\mathbf{v}_j,\mathbf{a}_i)/\tau}}\right], \tag{3.6}$$

where $\tau = 0.07$ is a temperature hyper-parameter and $\texttt{sim}(\cdot,\cdot)$ denotes the cosine similarity. Both terms in Equation 3.6 help bring $\mathbf{v}_i$ and $\mathbf{a}_i$ (i.e. the positives) together. The key difference is whether the negative set is composed of audio representations $\mathbf{a}_j$ or visual representations $\mathbf{v}_j$ where $j \neq i$

For readability of Equation 3.6, we once again absorb the notation for the audio and visual projection MLP heads ($h_A^{\texttt{AVC}}$ and $h_V^{\texttt{AVC}}$) within $\texttt{sim}(\cdot,\cdot)$, and illustrate their usage in Figure 3.4. Figure 3.4 also shows that we apply the AVC loss twice to associate both the visual clips, one that is extracted slightly before and another clip after the moment of interaction t, to the corresponding audio.

$\mathcal{L}_{\texttt{AVC}}$ vs $\mathcal{L}_{\texttt{AStC}}$: An important thing to note here is that AVC differs from the proposed AStC task. AVC seeks to associate the audio a_t with the corresponding visual clips v_t, as opposed to the change in visual state $\Delta\mathbf{v}_t$. Whereas, AStC seeks to associate the audio a_t with the dynamic/temporal change in the visual state. As a result, visual representations learned through AVC are biased towards static concepts, while those learned through AStC are more sensitive to dynamic concepts. Both types of representations are useful for egocentric tasks so that the representations have a spatial object understanding as well as a state-aware understanding.

3.2.4 Training

We learn the audio-visual representation encoders f_A and f_V and train them to minimize both AVC and AStC losses, such that the final objective function is as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{\texttt{AVC}} + (1 - \alpha) \mathcal{L}_{\texttt{AStC}} \tag{3.7}$$

where α is a weight hyper-parameter between the two terms.

Chapter 4

Experiments

4.1 Datasets

We evaluate on two egocentric datasets (Figure 4.1): **EPIC-Kitchens-100** [16] which contains 100 hours of activities in the kitchen and **Ego4D** [29] that contains 3670 hours of egocentric video covering daily activities in the home, workplace, social settings, etc. For experiments on Ego4D, we use all videos from the Forecasting and Hand-Object interaction subsets.

4.2 Implementation Details

We follow prior work on audio-visual correspondence [50], and use R(2+1)D video encoder [69] with depth 18 and a 10-layer 2D CNN as the audio encoder. We extract two short video clips around moments of interaction with a duration of 0.5 seconds and a frame rate of 16 FPS. The two video clips are separated by a gap of 0.2 seconds. Video clips are augmented by random resizing, cropping, and horizontal flipping resulting in clips of 8 frames at a resolution of 112 × 112. For audio, we extract an audio clip of 2 seconds at 44.1kHz and downsample them to 16kHz. Given the stereo audio, we average the two waveforms to get mono audio and then convert the mono signal to a log mel spectrogram with 80 frequency bands and 128 temporal frames. We train our algorithm with stochastic gradient descent for 100 epochs with

Figure 4.1: (Left) **EPIC-Kitchens-100**: Consists of 100 hours of activities in the kitchen, (Right) **Ego4D**: Contains 3670 hours of egocentric video covering daily activities in the home, workplace, social settings.

a batch size of 128 trained over 4 GTX 1080 Ti GPUs, a learning rate of 0.005 and a momentum of 0.9. For Ego4D, we keep the same parameters except for a batch size of 512 trained over 8 RTX 2080 Ti GPUs with a learning rate of 0.05. The loss terms in Equation 3.7 are equally weighted with an $\alpha = 0.5$.

4.3 Baselines

We consider the baselines which explore audio-visual modalities. *AVID* [50] and *XDC* [3] are two state-of-the-art models. AVID explores audio visual representation learning and alignment via contrastive learning, whereas XDC explores clustering based methods to achieve the same. AVID is pre-trained on 2M audio-visual pairs from AudioSet [26] that only leverages audio-visual correspondence. For our method, we initialize the model weights from AVID before training on moments of interaction to minimize both $\mathcal{L}_{\texttt{AVC}}$ and state change loss $\mathcal{L}_{\texttt{AStC}}$. We also compare our approach with the fully supervised methods presented in Ego4D [29].

4.4 Ablations

Random represents an untrained (randomly initialized) model, **RepLAI from scratch** is our method trained without AVID initialization, **RepLAI w/o $\mathcal{L}_{\texttt{AVC}}$** (Section 3.2.2, Section 3.2.3) is our method with only audible state change loss and

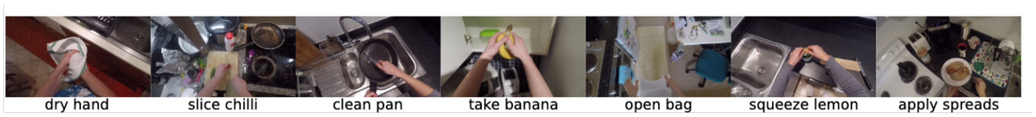

Figure 4.2: Downstream Tasks: Action Recognition (AR) of verb and noun

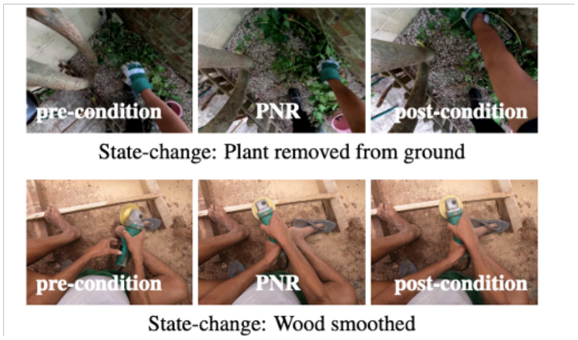
pre-condition
PNR
post-condition
State-change: Plant removed from ground
pre-condition
PNR
post-condition
State-change: Wood smoothed

Figure 4.3: Downstream Tasks: State Change Classification (StCC) and Point-of-no-return (PNR) temporal localization error

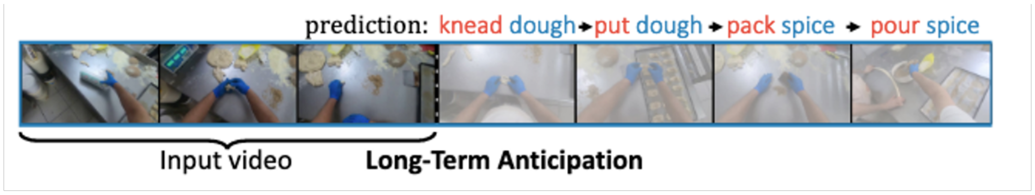
prediction: knead dough → put dough → pack spice → pour spice
Input video
Long-Term Anticipation

Figure 4.4: Downstream Tasks: Long-term action anticipation (LTA)

without the audio-visual correspondence loss, **RepLAI w/o $\mathcal{L}_{\texttt{AStC}}$** is our method with only audio-visual correspondence loss and without audible state change loss, **RepLAI w/o MoI** is our method trained on random moments in time and without a moment of interaction (MoI).

4.5 Downstream Tasks

After training our model on the self-supervised losses, we evaluate the representations on a range of egocentric downstream tasks (Figures 4.4, 4.3, 4.4). For all the downstream tasks, we follow the standard procedure and append a task specific decoder to the backbone model and train the decoder on a small annotated dataset. We discuss the downstream tasks below:

1. *Video action recognition (AR) on EPIC-Kitchens-100 and Ego4D* (Figure 4.2): Given a short video clip, the task is to classify the 'verb' and 'noun' of the action taking place in the video. We use two separate linear classifiers trained for this task and report the top-1 and top-5 accuracy on EPIC-Kitchens-100 [16] (Table 4.1) and Ego4D [29] (Table 4.2). We also evaluate on the unseen participants, head classes, and tail classes of EPIC-Kitchens-100 in Table 4.3. This task is helpful in assessing the spatio-temporal representations learned by the model in differentiating among different verbs and nouns.

2. *State change classification (StCC) on Ego4D* (Figure 4.3): Given a video clip, the task is to classify if an object undergoes a state change or not and is designed as a binary classification task. The video clip is encoded by f_V and a state change classification head is used which performs global average pooling on the entire feature tensor and is followed by a classification layer. For this task, we use the metric, State Change Classification Accuracy (%), and report it in Table 4.2. This is an ideal task for our model as it evaluates if the model is able to learn state-aware representations and identify the temporal change happening in the state of an object in an action.

3. *Long-term action anticipation (LTA) on Ego4D* (Figure 4.4): Given a video, the task is to predict the camera wearer's future sequence of actions. The model takes as input 4 consecutive clips of 2 seconds, which are encoded using our visual backbone f_V. Following [29], the representations are concatenated and given to 20 separate linear classification heads to predict the future 20 actions. We measure the performance using the edit distance metric ED@(Z=20) proposed in the work [29].[1] This task is helpful to evaluate if the representations can be used for long-horizon planning where the actions can change and may be of arbitrary duration. Results are reported in Table 4.2.

4. *Point-of-no-return (PNR) temporal localization error* (Figure 4.3): Given a video clip of a state change, the network has to estimate the time at which a state change begins. Specifically, the model tries to estimate the keyframe within the action video clip that contains the point-of-no-return (the time when

[1] Edit distance measures the minimum number of operations required to convert the predicted sequence of actions to ground truth. To account for multi-modality of future actions, it also allows the model to make $Z = 20$ predictions, and only accounts for the best prediction.

Method	$\mathcal{L}_{\texttt{AVC}}$	$\mathcal{L}_{\texttt{AStC}}$	**MoI Sampling**	**AVC Pretraining** [50]	**Top1 Acc ↑**		**Top5 Acc ↑**	
					Verb	Noun	Verb	Noun
(1) Random					20.38	4.96	64.75	19.83
(2) XDC [3]					24.46	6.75	68.04	22.71
(3) AVID [50]				✓	26.62	9.00	69.79	25.50
(4) RepLAI w/o `AVC`		✓	✓	✓	29.92	10.46	70.58	29.00
(5) RepLAI w/o `AStC`	✓		✓	✓	29.29	9.67	73.33	29.54
(6) RepLAI w/o MoI	✓	✓		✓	28.71	8.33	73.17	27.29
(7) RepLAI (scratch)	✓	✓	✓		25.75	8.12	71.25	27.29
(8) RepLAI	✓	✓	✓	✓	**31.71**	**11.25**	**73.54**	**30.54**

Table 4.1: Action recognition on EPIC-Kitchens-100. Top1 and top5 accuracy (%) is reported. ↑: Higher is better.

					StCC	**AR**		**LTA**		**PNR**
Method	$\mathcal{L}_{\texttt{AVC}}$	$\mathcal{L}_{\texttt{AStC}}$	**MoI**	**AVC Pretraining** [50]	**Acc ↑**	**Top1 Acc ↑**		**ED@(Z=20) ↓**		**Err ↓**
						Verb	Noun	Verb	Noun	
(S1) I3D-ResNet-50 [10, 29]	NA	NA	NA	NA	68.70	-	-	-	-	0.739
(S2) SlowFast [23, 29]	NA	NA	NA	NA	-	-	-	0.747	0.808	
(S3) MViT [21, 29]	NA	NA	NA	NA	-	-	-	0.707	0.901	
(1) Random					51.80	17.4	7.7	0.831	0.936	0.827
(2) XDC [3]					58.90	17.90	8.70	0.823	0.928	0.820
(3) AVID [50]				✓	61.11	18.3	10.7	0.811	0.919	0.814
(4) RepLAI w/o `AVC`		✓	✓	✓	64.00	20.3	12.4	0.781	0.854	0.792
(5) RepLAI w/o `AStC`	✓		✓	✓	63.60	21.1	13.5	0.774	0.853	0.795
(6) RepLAI w/o MoI	✓	✓		✓	62.90	19.8	11.2	0.792	0.868	0.801
(7) RepLAI (scratch)	✓	✓	✓		66.20	22.2	14.1	0.760	0.840	0.775
(8) RepLAI	✓	✓	✓	✓	**66.30**	**22.5**	**14.7**	**0.755**	**0.834**	**0.772**

Table 4.2: Performance on several downstream tasks on Ego4D. StCC: State Change Classification (%). AR: Action Recognition (%). LTA: Long-term action anticipation. ↑: Higher is better. ↓: Lower is better.

the state change begins). We do this by training a fully-connected head applied to each frame's representation in order to identify the timestamp at which the state of an object changes. We measure the performance using temporal localization error (seconds) in Table 4.2.

4.6 Discussion of Results

From Tables 4.1 and 4.2, we can observe that our method outperforms all other methods across all downstream tasks. We attribute this performance gain as we train the model to focus on interactions, both by detecting when they occur and by learning state-aware representations that are sensitive to interactions. We discuss the results in more detail for closer analysis and to draw some insights.

	Unseen Participants				Tail Classes				Head Classes			
	Top1 Acc ↑		**Top5 Acc ↑**		**Top1 Acc ↑**		**Top5 Acc ↑**		**Top1 Acc ↑**		**Top5 Acc ↑**	
Methods	Verb	Noun	Verb	Noun	Verb	Noun	Verb	Noun	Verb	Noun	Verb	Noun
XDC [3]	24.29	6.96	67.79	23.00	15.89	4.17	44.92	9.77	24.78	6.95	72.28	24.74
AVID [50]	26.17	8.67	68.75	24.12	16.80	4.43	47.14	12.89	27.95	9.82	73.20	28.43
RepLAI w/o AVC	28.67	9.38	72.04	27.88	18.49	5.21	47.79	12.63	30.59	10.21	73.33	30.90
RepLAI w/o MoI	27.71	7.92	72.08	26.88	16.80	4.04	49.74	12.76	29.36	10.65	76.33	30.41
RepLAI	**31.58**	**10.17**	**73.46**	**29.96**	**20.05**	**6.12**	**52.08**	**16.54**	**33.41**	**11.58**	**77.77**	**34.33**

Table 4.3: Video action recognition (AR) accuracy (%) on EPIC-Kitchens-100 for unseen participants, head classes, and tail classes. Top1 and top5 accuracy (%) is reported. ↑: Higher is better.

RepLAI enhances large-scale AVC driven approaches. Prior work on self-supervised audio-visual learning has shown strong audio-visual representations for action recognition [49, 50]. In our work, we try to explore how useful these representations are for egocentric tasks and what are their limitations. To do this, we compare our model trained from scratch, *RepLAI (Scratch)*, with our model using the weights from AVID [50] as initialization for both the visual and audio encoders. We also compare our method to standalone AVID and XDC i.e. without further self-supervised training. Comparing rows (2), (3), and (8) in Table 4.1 and Table 4.2, it is clear that our method enhances large-scale AVC pre-training by significant margins, leading to absolute improvements of 5% in top-1 verb accuracy on EPIC-Kitchens-100, 4.2% on Ego4D, 5.2% increase in state-change classification accuracy, 5.6% reduction on the edit distance for long-term anticipation compared to AVID, and 4.2% improvement in point-of-no-return localization error. Comparing rows (7) and (8), we also see that large-scale AVID pre-training enhances the representations learned by our method on EPIC-Kitchens-100 significantly but only marginally on Ego-4D. This is likely due to the significantly large diversity of scenes in Ego4D. Thus, while relying on large-scale audio-visual pre-training (as with AVID) can help avoid overfitting on smaller egocentric datasets, this is less critical when training on larger and more diverse data. This also shows that our method is able to leverage the large-scale unlabeled dataset and does not require initial pretraining given enough large-scale data.

Detecting moments of interaction (MoI) helps representation learning. We hypothesize that to learn good representations for egocentric data of daily activities, self-supervised learning should focus on moments in time when interactions occur. To assess whether our audio-driven MoI detection (Section 3.2.1) algorithm helps representation learning, we compare *RepLAI* with an ablated version, *RepLAI w/o MoI*, where the model is trained on audio-visual clips extracted at *random* from the untrimmed videos. As can be seen by comparing rows (6) and (8) in Table 4.1 and Table 4.2, sampling clips around MoI leads to significantly better representations for all egocentric downstream tasks that we study.

Moreover, even though *RepLAI w/o MoI* trains with `AStC` (Section 3.2.2), it is unable to fully leverage the state change objective function without the information of moments of interactions which leads to worse performance. Moments of interaction are helpful in giving a signal for state changes whereas random moments can consist of no activity or no interaction segments of data and are less likely to consist of state changes. This suggests that, an explicit state change objective function and sampling video clips around moments of interactions (which are likely to be aligned with the actual state changes) together provide an information-rich feedback to our model in better understanding of how the state changes by an interaction and how the actions transition over time. These results also clearly show that the proposed MoI detection procedure is able to find moments in time that are especially useful for learning representations of daily activities. We emphasize the simplicity and effectiveness of our audio-driven detector, which shows how informative audio can be when searching for moments of interaction. In the future, we believe that learning-based approaches could further enhance MoI detection, and further improve the learned audio-visual representations. We also show several qualitative examples of detected MoI in the supplement.

`AVC` and `AStC` are complementary. To assess the impact of both terms in Equation 3.7, we evaluate our method trained without $\mathcal{L}_{\texttt{AVC}}$ (Section 3.2.3) and without $\mathcal{L}_{\texttt{AStC}}$ (Section 3.2.2). Comparing rows (4), (5) to row (2) and row (3) in Table 4.1 and Table 4.2 shows that each term enhances the representations obtained through large-scale audio-visual pre-training (AVID) as compared to the baselines. Furthermore, comparing the ablated models in rows (4) and (5) to the full model in row (8) shows that these two terms are complementary to each other. This is because the `AVC` and `AStC` tasks encourage the learning of representations with different characteristics.

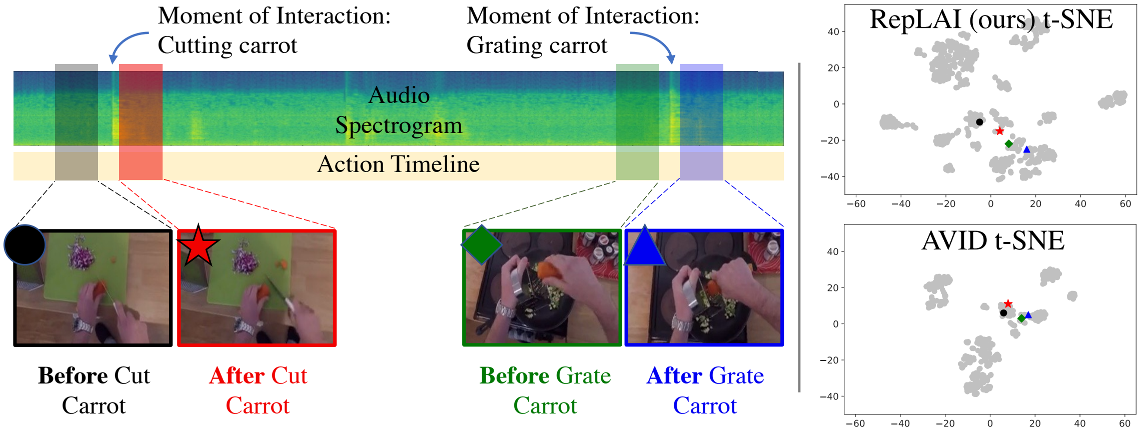

Figure 4.5: t-SNE visualization of the feature representations learned by *RepLAI* and *AVID* for a video consisting of fine-grained actions over time. For a simpler visualization, we consider all the videos belonging to a single participant. A larger spread in the t-SNE of *RepLAI* indicates more distinct state-aware representations.

`AVC` focuses on learning visual representations that are informative of what kind of sounding objects are present in the video, while `AStC` forces the model to differentiate between visual representations that occur before and after state change interactions.

RepLAI encourages state-aware representation learning. To study the representations learned by our approach for different states, we generate a t-SNE plot [70] for *RepLAI* and *AVID* as shown in Figure 4.5. For generating a simpler visualization, a small dataset is prepared consisting of all the videos corresponding to a single participant, *P01*, in EPIC-Kitchens-100 and split into clips of 0.5s. We can observe that there is a larger spread in the t-SNE plot for *RepLAI* than *AVID*. A larger spread indicates that the representations of the various states are significantly different from each other and form more distant clusters as shown by *RepLAI*. Whereas, if the state representations are similar to each other, they are clustered together and show lesser spread as shown by *AVID*. MoI are the key moments of interactions with an object in an environment where the state is changing. *AVID* has no such information about the key moments and also does not have an explicit state change objective function. Therefore, it is unable to discriminate between the *before* and *after* state of an action and has less effective state-aware information in its representations.

RepLAI representation are more generalizable and robust to long-tail. To assess RepLAI in a scenario with domain shift, we evaluate on unseen participants that were fully excluded from the pre-training of RepLAI. Table 4.3 shows that RepLAI significantly outperforms baselines and ablations, indicating that representation learning by our model provides much better generalization. Moreover, the verb and noun classes in EPIC-Kitchens-100 exhibit a long-tailed distribution. When further compared on head and tail classes separately in Table 4.3, we can observe that RepLAI outperforms all other methods highlighting its higher robustness on a long-tailed distribution.

Self-supervised vs supervised representation learning Table 4.2 also compares RepLAI to fully supervised methods introduced in Ego4D [29] (rows S1, S2 and S3). We can observe that RepLAI can also perform competitively to the fully supervised approaches when we have access to larger and more diverse data. With a further focus on SSL for untrimmed datasets, SSL methods will be able to match supervised approaches, and our work takes a step toward it.

4.7 Analysis of the types of interaction and potential failure modes

To provide further insights into the generalization ability of the proposed method, we conduct an experiment to assess how discriminative the learned representations are for different types of interactions. For this experiment, we first categorize the activities based on the nature of the transition:

- **T1**: irreversible interactions, backward transition highly unlikely (e.g., cut vegetables)
- **T2**: reversible interactions, backward transition occurs often (e.g., open/close fridge)
- **T3**: interactions with no transition direction (e.g., stirring)

`AStC` learns from both T1 and T2 interactions, as they are associated with visual state changes. Although T1 interactions are never seen in reverse order, the model still benefits from knowing the correct order, as this leads to more state-aware representations. As for T3 type interactions, they can be a failure mode of the `AStC` objective, if they cause no change in the visual state of the environment.

	Mean Average Precision			Norm of Visual State Change			Average Similarity		
Method	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**	**T1**	**T2**	**T3**
AVID	34.5	22.8	10.64	34.5	22.8	10.64	34.5	22.8	10.64
RepLAI	46.22	29.47	14.78	46.22	29.47	14.78	46.22	29.47	14.78

Table 4.4: Assessment of generalization ability of our method

4.7.1 Generalization and state change identifiability

To analyze how RepLAI representations behave for different types of interaction, we show several metrics in Table 4.4. We computed the mean average precision, after training a linear classifier for action recognition on Epic-Kitchens. The results indicate the RepLAI performs significantly better than the finetuned AVID baseline across all categories of transition/direction, showing that RepLAI (which includes both AVC and AStC) is generic enough to enhance representations for all types of interactions.

We also observed that MoI detection helps to find timestamps that have more perceptible visual state change (even for T3 type interactions) in Table 4.4. To see this, we computed the norm of the visual state change $||f_v(v_{t+\delta}) - f_v(v_{t-\delta})||$ around MoIs and around randomly chosen timestamps. We also measured how well the `AStC` loss learns the association between the audio and the visual state change in the forward direction. Specifically, we calculated the average similarity $sim(\Delta v_t^{frwd}, a_t)$ within each of the three categories (T1, T2, T3). Table 4.4 shows a comparison of this forward association score between RepLAI and the AVID baseline. As expected, RepLAI learns better associations between the audio and visual state changes than AVID. More importantly, despite being harder to identify, RepLAI still performs relatively well among T3 type interactions. This shows that, even for actions like washing and stirring, there are still slight visual state changes that the model can learn.

Chapter 5

Conclusions and Future Work

In this work, we propose a multi-modal audio-visual contrastive-learning based method for learning representations of egocentric videos of daily activities. We address two important challenges in order to learn strong representations for this domain. First, a model should focus learning on moments of interaction (MoI). Since these moments only occur sporadically in untrimmed videos, we show that MoI detection is an important component of representation learning in untrimmed datasets. This solves the issue of training on non-activity segments in long, untrimmed videos. Second, learning should focus on the natural aspects that arise from interactions, *i.e.*, the sharp audio signal corresponding to an interaction and changes in the state of an environment caused by agents interacting with the world. In particular, by seeking to identify visible state changes from the audio alone, we can learn representations that are potentially more aware of the state of the environment and hence, particularly useful for egocentric downstream tasks.

As a future work, we believe there are two components that can make our approach more robust and attend to the variable situations of the real-world. First, since noisy audio can result in false moments of interaction, the method can be made robust to noisy audio. Currently, we remove the background noise by normalizing the spectrogram across the frequencies, however, more robust approaches can be implemented to address the noisy audio of real-world and create more accurate moments of interaction. Since spectrogram is more sensitive to noise, a simple MLP-based approach can be used which can detect the moments of interaction and then in turn, its predictions can be used to train the network with the losses. Second, while the audible state change loss does give a signal to the network about the temporal state change direction, there is some room for improvement for the approach to perform well on actions that have no sense of direction. While these are some of the future possibilities to improve the model, our future vision is to create a model that can attend to any real-world scenarios by first detecting the moments of interaction accurately even in noisy scenarios and second, understanding the nature of action directions. Such a model can then be used to understand hand-object interactions in our daily activities, which can ultimately benefit robot manipulation tasks of interacting with an object.

Broader Impact: Deep learning models are generally capable of learning (and sometimes even amplifying) biases existing in datasets. While several steps have been taken in datasets like Ego4D to increase geographical diversity, we would like to encourage careful consideration of ethical implications when deploying these models. While public datasets are essential to make progress on how to represent visual egocentric data, premature deployment of our models is likely to have negative societal impact, as we did not check for the presence or absence of such biases.

Appendix A

Appendix

A.1 Additional downstream evaluation tasks

We evaluated all models on the audio representations of our model on two downstream tasks, state change classification and action recognition.

Action Recognition (AR) w/ audio: For this task, video embeddings from f_V and audio embedding from f_A are concatenated together and passed through two separate linear classifiers to classify the 'verb' and 'noun' of the action occurring in the video clip. Performance is measured using top-1 accuracy (%).

State change classification (StCC) w/ audio: For this task, we concatenate the representations from both video and audio modalities. Using these concatenated representations as input, the occurrence of state change is then predicted by training a binary linear classifier. We then measure the performance using state-change classification accuracy (%).

Incorporating audio modality helps with performance: Additionally, by comparing Table A.1 with Table 4.2 we observe that the performance on state change classification (StCC) and action recognition (AR) improves by incorporating the

					StCC w/ Audio	AR w/ Audio	
Method	$\mathcal{L}_{\texttt{AVC}}$	$\mathcal{L}_{\texttt{AStC}}$	**MoI**	**AVC Pretraining** [50]	**Acc** ↑	**Top1 Acc** ↑	
						Verb	Noun
(1) Random					52.90	18.90	9.50
(2) XDC [3]					57.70	19.10	10.20
(3) AVID [50]				✓	61.30	19.80	12.30
(4) RepLAI w/o `AVC`		✓	✓	✓	64.60	22.70	14.00
(5) RepLAI w/o `AStC`	✓		✓	✓	64.40	21.40	13.00
(6) RepLAI w/o MoI	✓	✓		✓	64.10	20.80	11.70
(7) RepLAI (scratch)	✓	✓	✓		66.30	22.50	15.00
(8) RepLAI	✓	✓	✓	✓	**66.80**	**23.10**	**15.80**

Table A.1: Performance on several downstream tasks on Ego4D. StCC w/ Audio: State Change Classification (%), AR w/ Audio: Action Recognition (%). ↑: Higher is better.

audio modality which shows the usefulness of audio representations. The gain in incorporating audio can be seen across all models, but is more significant on action recognition.

A.1.1 Detection of moments of interaction

In the experiments section, we showed that MoI detection improves representation quality. We evaluated the utility of moments of interaction (MoI) through their impact on representation quality and performance on multiple downstream tasks. Particularly, comparing rows (6) and (8) of Table 4.1 and Table 4.2 demonstrates that sampling training clips around MoIs improve representation quality and transfer.

We believe that MoI detection is especially useful for finding moments in the video with more perceptible visual state changes. We validate this by computing the norm of the difference between the before and after visual state for a detected MoI (averaged over all detected MoIs). A higher visual state change norm indicates that the model is able to detect locations in the video that have a significant and meaningful visual state change. From the Table A.2, we observe that the norm of visual state change around the detected MoIs is significantly higher than that around randomly picked locations. This validates that MoI is more effective in picking locations with relatively better visual state change. This, in turn, provides a richer signal to the model to learn better representations and provide stronger performance on downstream tasks.

Method	**visual state change** ↑
Random location	2.73
Moment of Interaction (MoI)	3.14

Table A.2: Evaluating the detected moments of interaction

A.2 Qualitative Analysis

A.2.1 Audio-visual correspondence analysis

We analyse the audio-visual correspondence learned by our method, *RepLAI* and compare it with *AVID* [50] by generating a t-SNE plot in Figure A.1. This correspondence is helpful in assessing how well the model is able to predict if a short video clip and an audio clip correspond with each other or not. For a simpler visualization, a small dataset consisting of a single participant in EPIC-KITCHENS-100 is taken and split into clips of 0.5s. Both the audio and video features are visualized in the same space in the t-SNE plot and represented by gray dots. A few examples are selected randomly and their visual representation as well their audio representation is shown in colors. It can be observed that the audio-visual representation dots are closer in *RepLAI* representing better audio-visual correspondence compared to *AVID*. This indicates that the `AStC` is helpful in enhancing the correspondence learned between the video and audio.

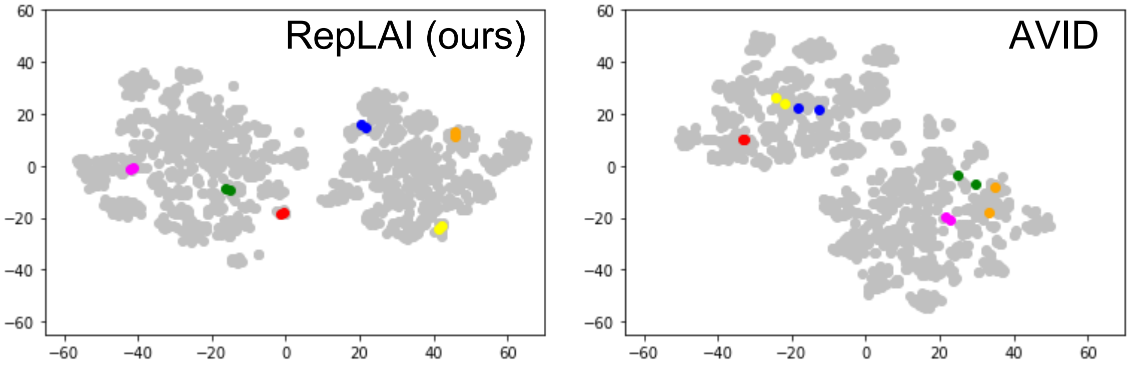

Figure A.1: t-SNE visualization of the audio-visual feature representations learned by *RepLAI* and *AVID*. For a simpler visualization, we consider all the videos of a single participant. The gray dots represent both the audio and visual features in the same space. Randomly 6 examples are chosen and their two dots are shown in colors representing the visual features and the audio features. Closer the two dots are, better the audio visual correspondence.

A.2.2 Detected MoIs

In this section, we visualize the moments of interaction detected with the help of spectrogram in several videos (Figure A.1, Figure A.2, and Figure A.3). While not perfect, we observe that sharp changes in the spectrogram energy correlate well with moments of interaction. Several of these moments are captured in Figure A.1, Figure A.2, and Figure A.3, such as opening drawers, putting down objects, cutting vegetables, etc.

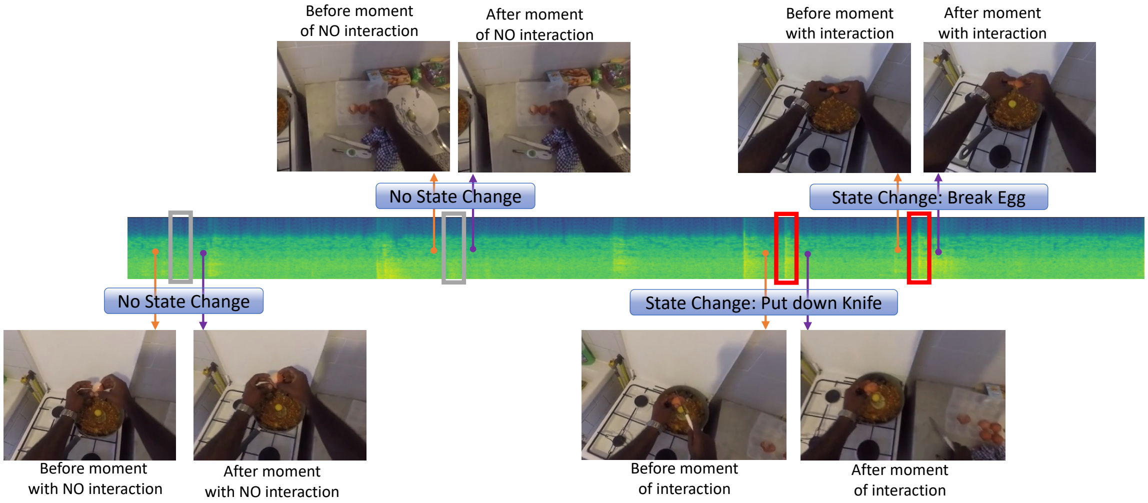

Figure A.1: The above visualization shows the spectrogram of a video containing the action of putting down knife and breaking egg. The gray indicate the random moments with no moment of interaction and red indicate the moments of interaction.

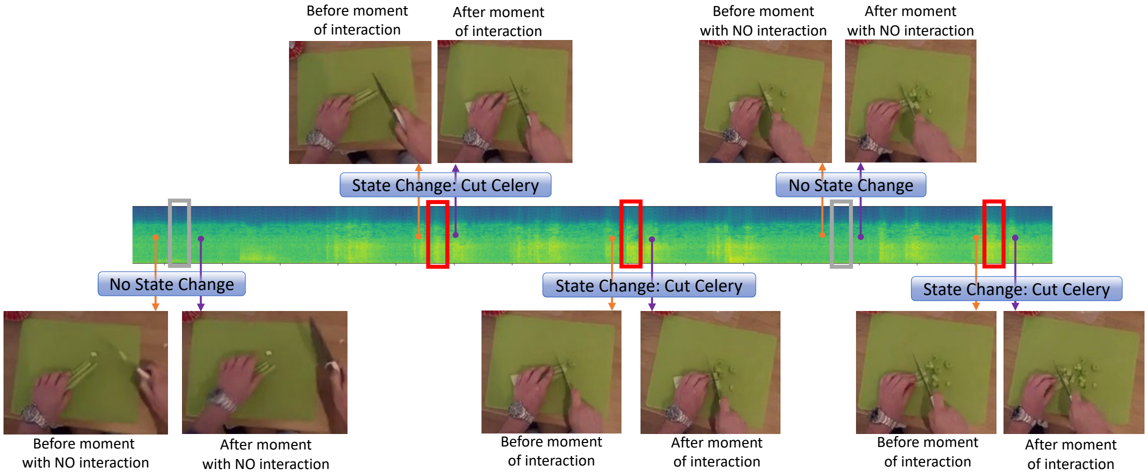

Figure A.2: The above visualization shows the spectrogram of a video containing the action of cutting celery. The gray indicate the random moments with no moment of interaction and red indicate the moments of interaction.

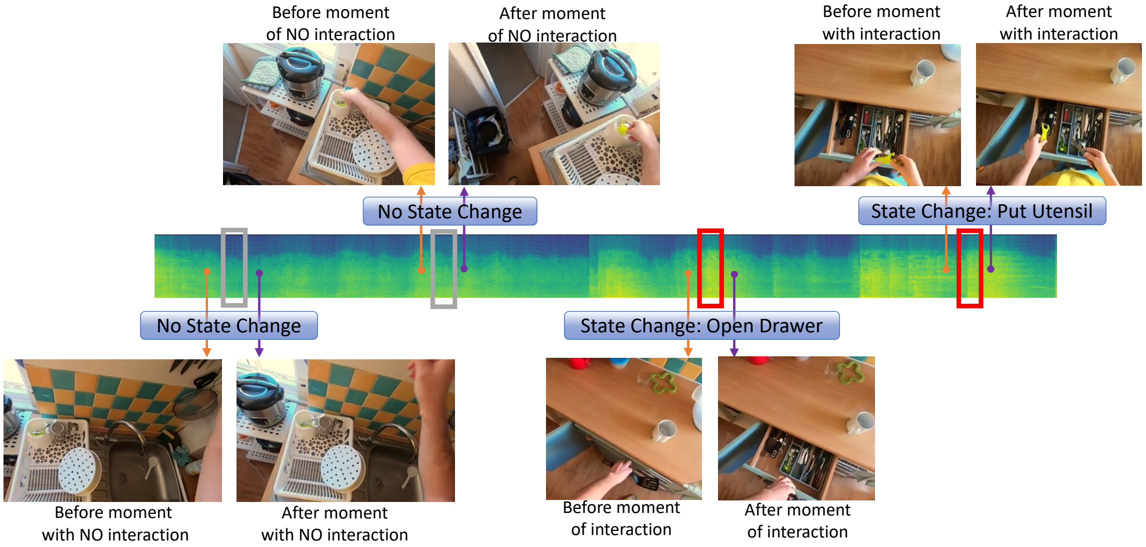

Figure A.3: The above visualization shows the spectrogram of a video containing the action of opening drawer and putting cutlery. The gray indicate the random moments with no moment of interaction and red indicate the moments of interaction.

Bibliography

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018.

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2 (6):7, 2020.

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020.

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 609–617, 2017.

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018.

[6] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020.

[7] Minjie Cai, Kris Kitani, and Yoichi Sato. Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes. *arXiv preprint arXiv:1807.08254*, 2018.

[8] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021.

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster

assignments. In *Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[11] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[13] Soo-Whan Chung, Joon Son Chung, and Hong-Goo Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969. IEEE, 2019.

[14] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, volume 2, page 3, 2014.

[15] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[16] 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. *arXiv preprint arXiv:2006.13256*, 2020.

[17] Virginia R de Sa. Learning classification with unlabeled data. In *Advances in Neural Information Processing Systems*, pages 112–119. Citeseer, 1994.

[18] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[19] Hazel Doughty and Cees GM Snoek. How do you do it? fine-grained action understanding with pseudo-adverbs. *arXiv preprint arXiv:2203.12344*, 2022.

[20] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1801–1810, 2019.

[21] 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.

[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, 2012.

[23] 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.

[24] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021.

[25] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020.

[26] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[27] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

[28] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021.

[29] 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 *CVPR*, 2022.

[30] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Guo, Mohammad Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*,

2020.

[31] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020.

[32] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pages 312–329. Springer, 2020.

[33] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021.

[34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[35] Hezhen Hu, Weilun Wang, Wengang Zhou, and Houqiang Li. Hand-object interaction image generation. *arXiv preprint arXiv:2211.15663*, 2022.

[36] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021.

[37] 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, 2019.

[38] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021.

[39] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8545–8552, 2019.

[40] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[41] 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*, pages 10138–10148, 2021.

[42] Kevin Lai, Liefeng Bo, and Dieter Fox. Unsupervised feature learning for 3d scene labeling. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3050–3057. IEEE, 2014.

[43] 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.

[44] Miao Liu, Siyu Tang, Yin Li, and James Rehg. Forecasting human object interaction: Joint prediction of motor attention and egocentric activity. 2019.

[45] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021.

[46] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022.

[47] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Proceedings of the European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[48] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33, 2020.

[49] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021.

[50] 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.

[51] 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, 2020.

[52] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019.

[53] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9890–9900, 2020.

[54] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Proceedings of the European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[55] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[56] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[57] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016.

[58] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020.

[59] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020.

[60] 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.

[61] Senthil Purushwalkam, Tian Ye, Saurabh Gupta, and Abhinav Gupta. Aligning videos in space and time. *arXiv preprint arXiv:2007.04515*, 2020.

[62] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021.

[63] Merey Ramazanova, Victor Escorcia, Fabian Caba Heilbron, Chen Zhao, and Bernard Ghanem. Owl (observe, watch, listen): Localizing actions in egocentric video via audiovisual temporal context. *arXiv preprint arXiv:2202.04947*, 2022.

[64] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021.

[65] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1961–1970, 2016.

[66] Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[67] Yu-Chuan Su and Kristen Grauman. Detecting engagement in egocentric video. In *European Conference on Computer Vision*, pages 454–471. Springer, 2016.

[68] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision*, pages 776–794. Springer, 2020.

[69] 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.

[70] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

[71] 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.

[72] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4006–4015, 2019.

[73] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *Proceedings of the European Conference on Computer Vision*, pages 504–521. Springer, 2020.

[74] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2794–2802, 2015.

[75] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2566–2576, 2019.

[76] Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR), June 2018.

[77] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019.

[78] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3895–3905, 2022.

[79] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021.

[80] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021.