LESS IS MORE
VIDEO UNDERSTANDING WITH REDUCED SUPERVISION

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTERS OF SCIENCE

Nishant Rai
May 2021

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a thesis for the degree of Masters of Science.

(Juan Carlos Niebles) Principal Advisor

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a thesis for the degree of Masters of Science.

(Ehsan Adeli) Secondary Advisor

Approved for the Stanford University Committee on Graduate Studies

Abstract

We have seen impressive progress in perception algorithms in recent years. Although we have seen a remarkable increase in performance as well as efficiency lately, their heavy dependence on annotated data limits applicability and hinders their deployment to newer tasks. We live in an era where data generation is happening at frightening speeds and scales, observing uploads of millions of pictures and videos daily. However, due to the complexity and manual labor involved in curating new datasets, such a massive reservoir of rich data is unable to be utilized by algorithms dependent on supervision.

Learning without supervision, i.e., unsupervised learning has long been treated as the holy grail of machine learning as it mimics the behaviour of humans who learn without the need for explicit supervision. Despite the presence of numerous approaches in this sub-field, their adoption was limited due to the reduced performance compared to their supervised counterparts. Recent approaches have explored reducing the need for supervision by utilizing implicit relationships present in unlabeled datasets. To this effect, research in multi view learning and self-supervised learning has led to impressive improvements. However, their intersection still remains largely unexplored. Although there has been rich research exploring the reduction of supervision in images, there is a dearth of similar approaches tackling videos which is arguably the future of perception. On the one hand, the additional temporal dimension leads to added complexity in videos, while on the other, it provides further implicit structure which can be used effectively.

Recently, self-supervision has proved itself to be a promising approach reducing the need for manual supervision drastically while maintaining competitiveness with other approaches. There has recently been a surge in interest for approaches utilizing self-supervised methods for visual representation learning. Recent advances in visual representation learning [9, 25] have demonstrated impressive performance even comparable to their supervised equivalents under specific settings. Continued research in this direction will enable future data-driven models to become increasingly more robust and performant using the vast amount of unlabeled data present in our world.

The idea of utilizing multiple modalities of information has been a well-established one with roots in human perception [11, 28]. It's argued that useful higher order semantics are present throughout different modalities and are consistent across them. At the same time, different modalities provide complementary information which can be utilized to aid learning in other modalities. Multi-view learning has been a popular direction [69, 80] utilizing these traits to improve representation quality. Recent approaches learn features

utilizing multiple modalities with the motivation that information shared across modalities has valuable semantic meaning.

My research goal in this thesis is to utilize the implicit structure in multi-modal videos to improve performance of perception algorithms as well as explore self-supervised approaches exploiting this implicit structure. It is argued that useful higher order semantics are present throughout different modalities and are consistent across them. At the same time, different modalities provide complementary information which can be utilized to aid learning.

In the beginning of this thesis, I explore the utility of multi-view learning and study its role in designing a self-supervised approach using multi-view associations. Recently, self-supervision has proved itself to be a promising approach reducing the need for manual supervision drastically while maintaining competitiveness with other approaches. I propose Cooperative Contrastive Learning, which utilizes multiple modalities of data to propose associations and leads to improved visual representations. Our main motivation is that each view sees a specific pattern, which can be useful to guide other modalities and improve representations.

Next, I present the Home Action Genome project with the goal of improving action understanding. Home Action Genome (HOMAGE) is a new benchmark for action recognition, that includes multi-modal synchronized videos from multiple viewpoints along with hierarchical action and atomic action labels. Actions in residential settings are challenging as we deal with long-term actions, interactions with objects, and frequent occlusions. Having multiple modalities and sensors to handle occlusions and scene graph information to capture object interaction allows us to tackle these complexities. I outline the impact having such rich modalities has on learning and potentially encouraging other lines of research.

After discussing the specifics of the Home Action Genome dataset, I discuss approaches which utilize the rich modalities present to build models which demonstrate improved performance. We also discuss the effectiveness of utilizing hierarchical labels and their effect on the overall robustness of our model. Together, we show the positive impact using cues from multiple camera views and modalities along with action compositions has on holistic performance.

Finally, I suggest multiple future applications of our dataset and proposed approaches. Overall, this thesis focuses on how utilizing multi-view datasets can both reduce the need for supervision while maintaining comparable performance as well as enabling existing models to improve performance and robustness. In conjunction, I hope this encourages research in this direction and brings us closer to the goal of learning without supervision.

Acknowledgments

First and foremost, I am extremely grateful to my advisors Juan Carlos Niebles and Ehsan Adeli for their invaluable advice, continuous support, and patience during my masters degree. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life.

Next, I would like to thank all my co-authors whose work is featured in this thesis (in alphabetical order): Adrien Gaidon, Haofeng Chen, Jingwei Ji, Kazuki Kozuka, Kuan-Hui Lee, Rishi Desai and Shun Ishizaka. Thank you to all members of the Stanford Vision and Learning Lab and the Video Understanding Group for the helpful discussions that contributed to this work.

Thank you to Stanford Computer Science Department, Stanford HAI and AWS, Toyota Research Institute and Panasonic for sponsoring parts of my research during my masters.

Finally, thank you to my family, friends, and loved ones for always being there for me. Thank you to Kopal Nihar for assisting me in balancing work and leisure along with helping me always stay sane. Thank you to my mom, Neetu Singh Rai, and dad, Om Prakash Rai, for always being there for me and encouraging me to follow my interests.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Learning without supervision has been treated as the holy grail of machine learning for a long time. It mimics the behaviour of humans who learn without the need for explicit supervision. Despite the presence of numerous approaches in this sub-field, their adoption was limited due to the reduced performance compared to their supervised counterparts. Recent approaches have explored utilizing implicit relationships present in unlabeled datasets. To this effect, research in multi view learning and self-supervised learning has led to impressive improvements. However, their intersection still remains largely unexplored. Although there has been rich research exploring the reduction of supervision in images, there is a dearth of similar approaches tackling videos which is arguably the future of perception. On the one hand, the additional temporal dimension leads to added complexity in videos, while on the other, it provides further implicit structure which can be used as supervision.

Self-supervision has been a promising development in reducing the amount of supervision needed to learn effectively. Recently, self-supervision has proved itself to be a promising approach reducing the need for manual supervision drastically while maintaining competitiveness with other approaches. There has recently been a surge in interest for approaches utilizing self-supervised methods for visual representation learning. Recent advances in visual representation learning [9, 25] have demonstrated impressive performance even comparable to their supervised equivalents under specific settings. Continued research in this direction will enable future data-driven models to become increasingly more robust and performant using the vast amount of unlabeled data present in our world. Yet another promising direction to learn improved representations is multi-view learning. The idea of utilizing multiple views of information has been a well-established one with roots in human perception [11, 28]. It's argued that useful higher order semantics are present throughout different views and are consistent across them. At the same time, different views provide complementary information which can be utilized to aid learning in other views. Recent contributions have explored the

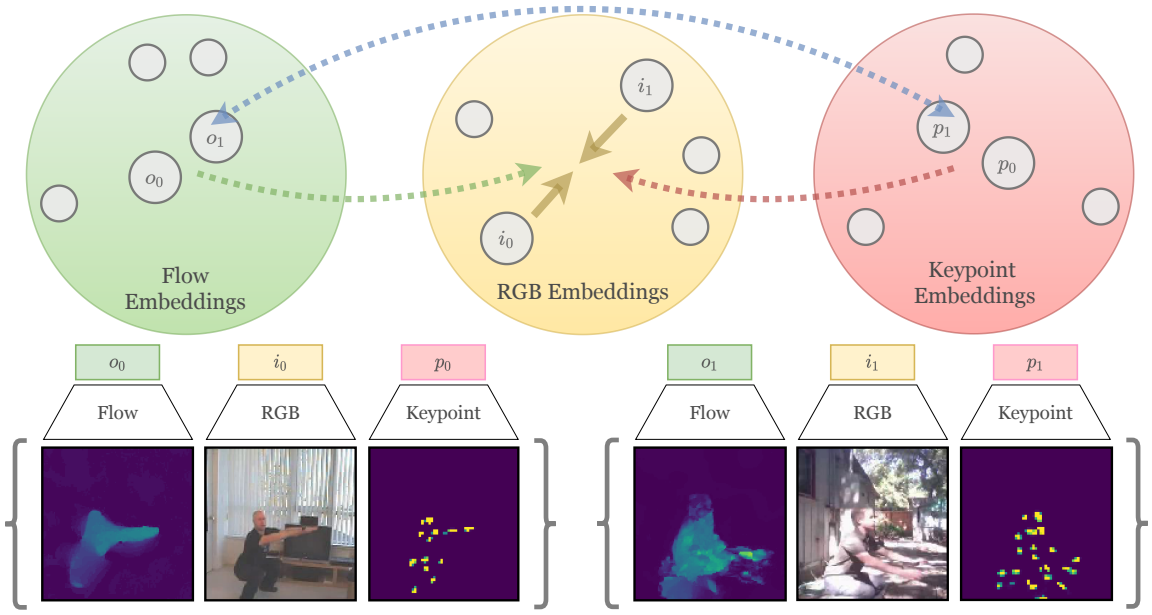


Figure 1.1: Given a pair of instances (e.g. people doing squats) and corresponding multiple views, features are computed using view-specific deep encoders f's. Different instances may have contrasting similarities in different views. For instance, V_0 (left) and V_1 (right) have similar optical-flow $o = f_{flow}$ and pose keypoints (keypoint) $p = f_{keypoint}$ features but their image $i = f_{rgb}$ features are far apart. CoCon leverages these inconsistencies by encouraging the distances in all views to become similar. High similarity of o_0, o_1 and p_0, p_1 nudges i_0, i_1 towards each other in the RGB space.

intersection of multi-view learning and self-supervised learning [69, 80] utilizing these traits to improve representation quality. Using both self-supervision as well as multi view information to aid action understanding is great combination to tackle the challenges posed by the problem.

Action understanding in videos is a critical task with various use cases and real-world applications, from robotics [47, 66] and human-computer interaction [68] to healthcare [23, 50] and elderly behavior monitoring [30, 51]. Despite the recent success of deep learning methods for image classification, complex and holistic action or event understanding remains an elusive task.

There are several challenges associated with the task of action understanding. The inherent variability in executing complex activities poses one of the most critical difficulties in building action understating models. To understand these challenges, it is essential to understand what actions are composed of. As opposed to bounding boxes in the object detection task, actions are composed of various parts spanned in space and time. For instance, the action of "laundry" involves multiple entities, e.g., humans, objects, and their relationships, and is composed of a number of atomic actions. Such partonomy of actions [5, 32, 89] both in space and time defines a hierarchical structure. Furthermore, to capture the variability in executing complex activities, understanding each part (e.g., body limbs, objects, or atomic actions) becomes crucial. Since actions happen in the 3D world, a holistic understanding of the world requires capturing the subtle movements or parts using

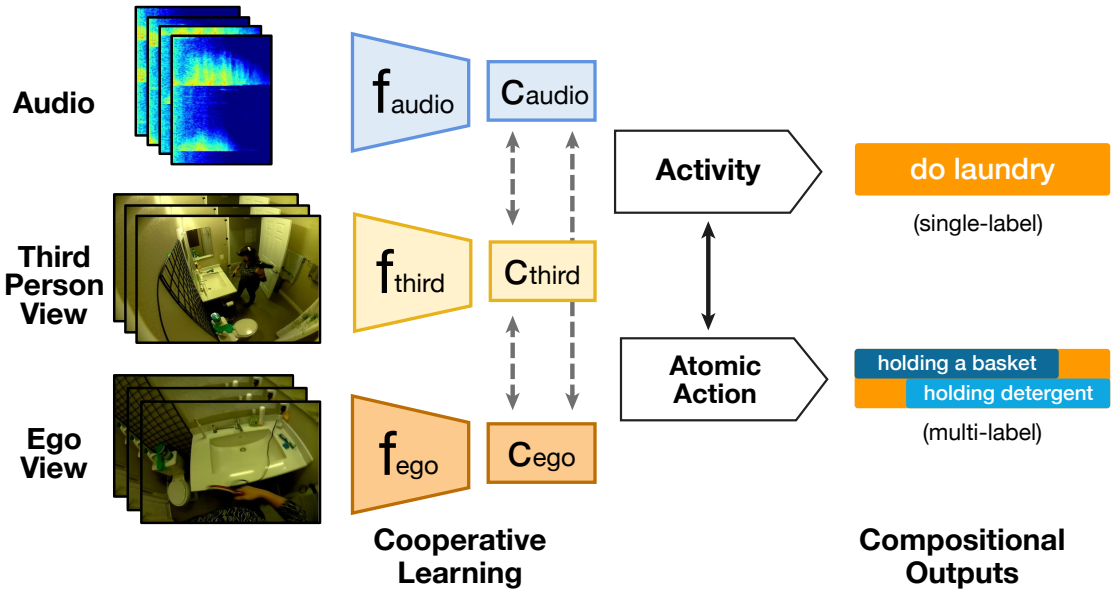


Figure 1.2: Given an activity instance (e.g., 'do laundry') and corresponding multiple views, we compute features using modality-specific deep encoders (f modules). Different modalities may capture different semantic information regarding the action. *Cooperatively* training all modalities together allows us to see improved performance. We utilize training using both video-level and atomic action labels to allow both the videos and atomic actions to benefit from the *compositional* interactions between the two. As discussed in the results, we see significantly improved performance when using the above components together.

multiple modalities (e.g., RGB and audio) and from multiple viewpoints.

Each of these challenges has previously been separately investigated using different datasets and advanced methods. For instance, numerous datasets were put together for generic action recognition and spatio-temporal localization in YouTube or broadcasting third-person videos, such as Kinetics [6], Charades [72], ActivityNet [14], UCF101 [76]. Other datasets such as EPIC Kitchens [10] were used for ego-centric action recognition. Action Genome [32] focused on using scene information in action recognition, while others [53] focused on hierarchical action modeling from events to low-level atomic actions. Several studies target learning from long instructional videos and release datasets [8, 55, 79, 91] for the same, exploring the partonomy of actions in long sequences. Others also focused on observing and recognizing actions from multiple views, such as LEMMA [34] and HumanEva [71]. In parallel, there have been numerous recent advances in contrastive and cooperative learning [9, 22] applied to multimodal and multi-view datasets as a self-supervised pre-training strategy to improve downstream recognition results. Despite all these advances, action understanding and generalizability of such models remains a challenging problem due to the complexities brought by their complicated nature and numerous object interactions. Multi-modal approaches [69, 74, 80] have shown superior performance in tackling such issues. However, there is still a need for a benchmark that unifies all these challenges and tasks. In this thesis, we discuss a dataset along with a novel method for hierarchical action recognition to tackle these problems.

Apart from multi-modal perception, self-supervised approaches allow us to further exploit explicit structure in our data to improve performance. Videos are a rich source for self-supervision, due to the inherent temporal consistency in neighboring frames. A natural approach to exploit this temporal structure is predicting future context as done in [20, 29, 49, 54]. Such approaches perform future prediction in mainly two ways: (1) predicting a reconstruction of future frames [49, 54, 77], (2) predicting features representing the future frames [20, 29]. If the goal is learning high-level semantic features for other downstream tasks, then complete reconstruction of frames is unnecessary. Inspired by developments in language modelling [57], recent work [83] propose losses that only focus on the latent embedding using frame-level context. One of the more recent approaches [20] propose utilizing spatio-temporal context to learn meaningful representations. Even though such developments have led to improved performance, the quality of the learned features is still lagging behind that of their supervised counterparts. In this thesis, we explore the intersection of multi-modal and self-supervised learning and propose a cooperative algorithm which allows us to learn improved video representations.

1.1.1 Thesis Outline

We discuss multiple contributions in this thesis. We first introduce an approach to utilize multi-modal cues to enable self-supervised video understanding resulting in a drastic reduction in the amount of supervision needed. We then introduce the Home Action Genome dataset, that includes multi-modal synchronized videos from multiple viewpoints along with hierarchical action and atomic action labels. Finally, we propose an approach allowing us to use the rich annotations provided in Home Action Genome, demonstrating the robustness and improved performance of our proposed model. The thesis proceeds as follows: In Chapter 2, we provide a detailed description of utilizing multi-view cues to propose a new self supervised algorithm. Chapter 3 discusses the Home Action Genome dataset in detail and touches upon the benefits of having such a richly annotated dataset. Chapter 4 presents an approach to utilize the multi-modal and multi-camera-view data along with the rich annotations provided in Home Action Genome. We discuss the performance benefits and generalizability of our new model through some quantitative and qualitative experiments. The thesis concludes in Chapter 5 where we briefly list potential future directions and applications of our research.

Further visualizations, code, dataset and additional information on the projects can be found here[1] and here[2].

1.1.2 Previously Published Papers

Most contributions in this thesis have first appeared as various publications. These publications are: [63] (Chapter 2) and [64] (Chapter 3, 4).

[1] https://github.com/nishantrai18/cocon
[2] http://homeactiongenome.org/

Chapter 2

Learning with Multiple Modalities

2.1 Introduction

There has recently been a surge in interest for approaches utilizing self-supervised methods for visual representation learning. Recent advances in visual representation learning have demonstrated impressive performance compared to their supervised counterparts [9, 25]. Fresh development in the video domain have attempted to make similar improvements [20, 29, 49, 69].

Videos are a rich source for self-supervision, due to the inherent temporal consistency in neighboring frames. A natural approach to exploit this temporal structure is predicting future context as done in [20, 29, 49, 54]. Such approaches perform future prediction in mainly two ways: (1) predicting a reconstruction of future frames [49, 54, 77], (2) predicting features representing the future frames [20, 29]. If the goal is learning high-level semantic features for other downstream tasks, then complete reconstruction of frames is unnecessary. Inspired by developments in language modelling [57], recent work [83] propose losses that only focus on the latent embedding using frame-level context. One of the more recent approaches [20] propose utilizing spatio-temporal context to learn meaningful representations. Even though such developments have led to improved performance, the quality of the learned features is still lagging behind that of their supervised counterparts.

Due to the lack of labels in self-supervised settings, it is impossible to make direct associations between different training instances. Instead, prior work has learned associations based on structure, either in the form of temporal [20, 38, 44, 52, 86] or spatial proximity [20, 35, 38, 59] of patches extracted from training images or videos. However, the contrastive losses utilized enforce similarity constraints between instances from same videos while pushing instances from other videos far away even if they represent the same semantic content. This inherent drawback forces learning of features with limited semantic knowledge and encourage performing low-level discrimination between different videos. Recent approaches suffer from this restriction leading to poor representations.

The idea of utilizing multiple modes of information has been a well-established one with roots in human

perception [11, 28]. It's argued that useful higher order semantics are present throughout different modes and are consistent across them. At the same time, different modes provide complementary information which can be utilized to aid learning in other modes. Multi-view learning has been a popular direction [69, 80] utilizing these traits to improve representation quality. Recent approaches learn features utilizing multiple modes with the motivation that information shared across modes has valuable semantic meaning. A majority of these approaches directly utilize core ideas such as contrastive learning [60] and mutual information maximization [4, 46, 88]. Although the fusion of modes leads to improved representations, such approaches also utilize contrastive losses, consequently suffering from the same drawback of low-level discrimination between similar instances.

We propose Cooperative Contrastive Learning (CoCon), which overcomes this shortcoming and leads to improved visual representations. Our main motivation is that each mode sees a specific pattern, which can be useful to guide other modes and improve representations. Our approach utilizes inter-view information to avoid the drawback of discriminating similar instances discussed earlier. To this end, each mode sees a different aspect of the videos, allowing it to suggest potentially similar instances to other modes. This allows us to infer implicit relationships between instances in a self-supervised multi-modal setting, something which we are the first to explore. These associations are then used in order to learn better representations for downstream applications such as video classification and action recognition. Fig. 5.2 shows an overview of CoCon. It is worth noting that although CoCon utilizes building blocks currently used in self-supervised representation learning, it is applicable to other tasks utilizing contrastive learning and be used in conjunction with other recently proposed methods. We provide more details about our approach in latter sections.

2.2 Related Work

2.2.1 Self-supervised Learning from images

Recent approaches have tackled image representation learning by exploiting color information [43, 90] and spatial relationships [59, 67], where relative positions between image patches are exploited as supervisory signals. Several approaches apply self-supervision to super-resolution [13, 36] or even to multi-task [12] and cross-domain [65] learning frameworks.

2.2.2 Self-supervised Learning from videos

Multiple approaches [20, 29, 49, 54, 77] perform self-supervision through 'predicting' future frames. However, the term 'predicting' is overloaded, as they do not directly predict and reconstruct frames but instead operate on latent representations. This ignores stochasticity of frame appearance, e.g., illumination changes, camera motion, appearance changes due to reflections and so on, allowing the model to focus on higher-order semantic features. Recent work [20, 80] utilize Noise Contrastive Estimation to perform prediction of the latent representations rather than the exact future frames, vastly improving performance. Yet, another class

of proxy tasks are based on temporal ordering of frames [56, 86]. Temporal coherence [31, 85] and 3D puzzle [38] were used as proxy loss to exploit spatio/temporal structures.

2.2.3 Multi-view learning

Multiple modes of videos are rich sources of information for self-supervised learning [69, 80, 84]. Two stream networks for action recognition [74] have led to many competitive approaches, which demonstrate using even derivable modes such as optical flow helps improve performance considerably. There have been approaches [52, 69, 80, 84] utilizing diverse modes, sometimes derivable from one other, to learn better representations. However, these approaches utilize inter-view links by maximizing mutual information between them. Although this leads to improved performance, we believe the rich inter-view linkages can be utilized more effectively by utilizing them to uncover implicit relationships between instances.

2.2.4 Multi-View Self-supervised learning

Multiple recent approaches [3, 21, 22, 62] have tackled the challenge of multi-modal self-supervised learning achieving impressive performance. However, these approaches suffer from the same drawback of discriminating between similar instances, leaving potential to benefit from inter-sample relationships.

Most approaches above perform self-supervision using positive and negative pairs mined through structural constraints, e.g., temporal and spatial proximity. Although this results in representations that capture some degree of semantic information, it incorrectly leads to treating similar actions differently due to the inherent nature of their pair-mining. For instance, clip pairs in different videos are considered negatives, even if they represent the same action. We argue that utilizing different modes and inter-instance relationships to propose positive pairs to aid training can lead to improvement of all modes simultaneously.

Chapter 3

Cooperative Contrastive Learning

3.1 Introduction

In the previous chapter, we discussed how multi modal learning can be integrated with self-supervised learning to learn effective video representations. In this section, we take a deeper look at our proposed approach and formulate how we utilize multi modal cues to cooperatively improve learnt representations across modalities.

We propose Cooperative Contrastive Learning (CoCon), which overcomes this shortcoming and leads to improved visual representations. Our main motivation is that each view sees a specific pattern, which can be useful to guide other views and improve representations. Our approach utilizes inter-view information to avoid the drawback of discriminating similar instances discussed earlier. To this end, each view sees a different aspect of the videos, allowing it to suggest potentially similar instances to other views. This allows us to infer implicit relationships between instances in a self-supervised multi-view setting, something which we are the first to explore. These associations are then used in order to learn better representations for downstream applications such as video classification and action recognition. Fig. 5.2 shows an overview of CoCon. It is worth noting that although CoCon utilizes building blocks currently used in self-supervised representation learning, it is applicable to other tasks utilizing contrastive learning and be used in conjunction with other recently proposed methods.

We use 'freely' available views of the input such as RGB frames and Optical Flow. We also explore the benefit of using high-level inferred semantics as additional noisy views, such as human pose keypoints and segmentation masks generated using off-the-shelf models [87]. These views are not independent, as they can be derived from the original input images. However, they are complementary and lead to significant gains, demonstrating CoCon's effectiveness even with noisy related views. The extensible nature of our framework and the 'freely' available views used make it possible to use CoCon with any publicly available video dataset and other contrastive learning approaches.

3.2 Method

We describe cooperative contrastive learning (CoCon) and intuition behind our designs in this section.In the following sections, we build our framework borrowing the learning framework present in [20] which learns video representations through spatio-temporal contrastive losses. It should be noted that even though we use this particular self-supervised backbone in our experiments, our approach is not restricted by the choice of the underlying self-supervised task. CoCon can be used in conjunction with any other frameworks currently present and allow them to be extended to a multi-modal setting.

A video V is a sequence of T frames (not necessarily RGB images) with resolution $H \times W$ and C channels, $\{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_T\}$, where $\mathbf{i}_t \in \mathbb{R}^{H \times W \times C}$. Assume $T = N * K$, where N is the number of blocks and K denotes the number of frames per block. We partition a video clip V into N disjoint blocks $V = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_j \in \mathbb{R}^{K \times H \times W \times C}$ and a non-linear encoder $f(.)$ transforms each input block x_j into its latent representation $z_j = f(x_j)$. An aggregation function, $g(.)$ takes a sequence $\{z_1, z_2, \ldots, z_j\}$ as input and generates a context representation $c_j = g(z_1, z_2, \ldots, z_j)$. In our setup, $z_j \in \mathbb{R}^{H' \times W' \times D}$ and $c_j \in \mathbb{R}^D$. D represents the embedding size and H', W' represent down-sampled resolutions as different regions in z_j represent features for different spatial locations. We define $z'_j = Pool(z_j)$ where $z'_j \in \mathbb{R}^D$ and $c = F(V)$ where $F(.) = g(f(.))$.

Similar to [20], we create a prediction task involving predicting z of future blocks. For multiple modes, we define $c_v = F_v(V_v)$, where V_v, c_v and F_v represent the input, context feature and composite encoder for mode v respectively.

3.2.1 Contrastive Loss

Noise Contrastive Estimation (NCE) [18, 57, 60] constructs a binary classification task where a classifier is fed with real and noisy samples with the training objective being distinguishing them. Similar to [20, 60], we use an NCE loss over our feature embeddings described in Eq 3.1. $z_{i,k}$ represents the feature embedding for the i^{th} time-step and the k^{th} spatial location. Recall $z_j \in \mathbb{R}^{H' \times W' \times D}$ which preserves the spatial layout. We normalize $z_{i,k}$ to lie on the unit hypersphere. Eq 3.1 is a cross-entropy loss distinguishing one positive pair from all the negative pairs present in a video. We use temperature $\tau = 0.005$ in our experiments. In a batch setting with multiple video clips, it is possible to have more inter-clip negative pairs.

To extend this to multiple modes, we utilize different encoders ϕ_v for each mode v. We train these encoders by utilizing $\mathcal{L}_{cpc}$ for each of them independently, giving us, $\mathcal{L}_{cpc} = \sum_v \mathcal{L}^v_{cpc}$

$$\mathcal{L}_{cpc} = -\sum_{i,k} \left(\log \frac{\exp(\tilde{z}_{i,k} \cdot z_{i,k} \,/\, \tau)}{\sum_{j,m} \exp(\tilde{z}_{i,k} \cdot z_{j,m} \,/\, \tau)} \right) \tag{3.1}$$

Figure 3.1: Examples for each mode. From top to bottom - RGB, Flow, SegMasks and Poses. Note the prevalence of noise in a few samples, specially SegMasks; There are multiple other instances where Poses, SegMasks are noisy but have not been shown here.

3.2.2 Cooperative Multi-View Learning

Recent approaches [22, 69, 80] tackle multi-modal self-supervised learning by maximizing mutual information across modes. They involve using positive and negative pairs generated using structural constraints, e.g., spatio-temporal proximity in videos [20, 21, 69, 80]. Although such representations capture semantic content, they unintentionally encourage discriminating video clips containing semantically similar content due to the inherent nature of pair generation, i.e. video clips from different videos are negatives. We utilize inter-instance relationships to alleviate some of these issues.

We soften this constraint by indirectly deriving pair proposals using different modes. Such a co-operative scheme benefits all models as each individual mode gradually improves. Better models are able to generate better proposals, improving performance of all modes creating a positive feedback loop. Our belief is that significant semantic features should be universal across modes, therefore, potential incorrect proposals from one mode should cancel out through proposals from other modes.

We achieve the above by computing mode-specific distances and synchronizing them across all modes. We enforce a consistency loss between distances from each mode. Looking at it from another perspective, we are encouraging relationships between instances to be the same across modes i.e. similar pairs in one mode should be a similar pair in other modes as well. Treating this as inter-view graph regularization, we create a graph similarity matrix W_v of size $K \times K$, using some distance metric. We represent our distance metric by $\mathcal{D}(.)$. In our experiments, we use the cosine distance which translates to $W^v_{ab} = z_z \cdot z_b$.

Assume h^a_v denotes the representation for the v^{th} mode of instance a. In our experiments, we use $h = z'$ giving us block level features. Our resultant loss becomes the inconsistency between similarity matrices across modes. The resultant graph regularization loss becomes $\sum_{v_0,v_1} \|W^{v_0} - W^{v_1}\|$ which is simplified in Eq 3.2.

Building on top of our earlier intuition, in order to have sensible proposals, we need to have discriminative scores, i.e. we should have both positive ($\mathcal{D} \to 0$) and negative ($\mathcal{D} \to 1$) pairs. To promote well distributed distances, we utilize the hinge loss described in Eq 3.3.

$\mathcal{L}_{sim}$ is the hinge loss, where the first term pushes representations of the same instance in different modes closer; while the second term pushes different instances apart. Since the number of structural negative pairs are much larger than the positives, we introduce μ in order to balance the loss weights. We choose μ such that the first and second components contribute equally to the loss.

$$\mathcal{L}_{sync} = \sum_{v_0,v_1} \sum_{a,b} \Big(\mathcal{D}(h^a_{v_0}, h^b_{v_0}) - \mathcal{D}(h^a_{v_1}, h^b_{v_1}) \Big)^2 \tag{3.2}$$

$$\begin{aligned} \mathcal{L}_{sim} = & \sum_{v_0,v_1} \sum_{a} \mathcal{D}(h^a_{v_0}, h^a_{v_1}) \\ & + \mu \sum_{a \neq b} max\big(0, 1 - \mathcal{D}(h^a_{v_0}, h^b_{v_1})\big) \end{aligned} \tag{3.3}$$

View	Random	$\mathcal{L}_{cpc}$	$\mathcal{L}^{cpc}_{sim}$	$\mathcal{L}^{cpc}_{sync}$	$\mathcal{L}_{cocon}$
RGB	46.7	63.7	66.0	62.7	**67.8**
Flow	65.3	69.8	71.4	69.2	**72.5**

Table 3.1: Impact of losses on performance of models when jointly trained with RGB and Flow. CoCon i.e. $\mathcal{L}_{total}$ (67.8) comfortably improves performance over CPC i.e. $\mathcal{L}_{cpc}$ (63.7). $\mathcal{L}^x_y = \mathcal{L}_x + \lambda\mathcal{L}_y$ where $\lambda = 10.0$ for this experiment

Method	Pretrain	RGB		Flow[2]	
		UCF	HMDB	UCF	HMDB
Random		46.7	20.6	65.3	31.2
CPC	K400	68.6	35.5	69.8	40.8
CoCon	UCF	67.8	37.7	**72.5**	44.1
CoCon	K400	**72.1**	**46.5**	71.8	**44.2**

Table 3.2: Impact of pre-training comparison. CoCon demonstrates a consistent improvement in both RGB and Flow.

Note that $\mathcal{L}_{sim}$ entangles different modes together. An alternative would be defining such a loss individually for each mode. However, diversity is inherently encouraged through $\mathcal{L}_{cpc}$, and interactions between modes have the side-effect of increasing their mutual information (MI), which leads to improved performance [69, 80].

We combine the above losses to get our cooperative loss, $\mathcal{L}_{coop} = \mathcal{L}_{sync} + \alpha \cdot \mathcal{L}_{sim}$. We use $\alpha = 1.0$ for our experiments and observe roughly similar performance for different values of α. The overall loss of our model is given by $\mathcal{L}_{cocon} = \mathcal{L}_{cpc} + \lambda \cdot \mathcal{L}_{coop}$. $\mathcal{L}_{cpc}$ encourages our model to learn good features for each mode, while $\mathcal{L}_{coop}$ nudges it to learn higher-level features using all modes while respecting the similarity structure across them.

Method	RGB		Flow		PoseHM		SegMask	
	UCF	HMDB	UCF	HMDB	UCF	HMDB	UCF	HMDB
Random	46.7	20.6	65.3	31.2	51.7	33.0	42.7	26.3
CPC	63.7	33.1	71.2	44.6	56.4	42.0	53.7	32.8
CoCon	**71.0**	**39.0**	**74.5**	**45.4**	**58.7**	**42.6**	**55.8**	**34.0**

Table 3.3: Impact of co-training on modes. CoCon is jointly trained with four modalities (RGB, Flow, PoseHM, & SegMask).

Action Class	CoCon	CPC
PlayCello	PlaySitar, PlayTabla, PlayDhol	N/A
Skiing	Surfing, Skijet	Surfing
HammerThrow	BaseballPitch, ThrowDiscus, Shotput	N/A
BrushTeeth	ApplyLipstick, EyeMakeup, ShaveBeard	ApplyLipstick

Table 3.4: Nearest consistent semantic classes. Individually trained modes (*CPC*) do not have consistent neighbors across modes, leading to empty results (N/A) for 'PlayingCello' and 'HammerThrow'. While modes trained using CoCon show consistency across modes, leading to sensible relationships e.g. 'HammerThrow' related to other classes involving throwing.

# Views	RGB		Flow	
	UCF	HMDB	UCF	HMDB
2	67.8	37.7	72.5	44.1
4	**71.0**	**39.0**	**74.5**	**45.4**

Table 3.5: Impact of performance on varying modes. A consistent improvement can be seen with more modes despite the prevalent noise in PoseHM and SegMasks.

3.3 Experiments

The goal of our framework is to learn video representations which can be leveraged for video analysis tasks. Therefore, we perform experiments validating the quality of our representations. We measure downstream action classification to objectively measure model effectiveness and analyze impact of our designs through controlled ablation studies. We also conduct qualitative experiments to gain deeper insights into our approach. In this section, we briefly go over our experiment framework.

3.3.1 Overview

Datasets

Our approach is a self-supervised learning framework for any dataset with multiple modes. However, we discuss its relevance to video action classification in our experiments. We focus on human action datasets i.e. UCF101, HMDB51 and Kinetics400. UCF101 contains 13K videos spanning over 101 human action classes. HMDB51 contains 7K video clips mostly from movies for 51 classes. Kinetics-400 (K400) is a large video dataset with 306K video clips from 400 classes.

Views We utilize different modes in our experiments. For Kinetics-400, we learn encoders for RGB and Optical Flow. We use Farneback flow (FF) [15] instead of the commonly used TVL1-Flow as it is quicker to compute lowering our computation budget. Although FF leads to lower performance compared to TVL1, the essence of our claims remain unaffected. For UCF101 and HMDB51, we learn encoders for RGB, TVL1 Optical Flow, Pose Heatmaps (PoseHMs) and Human Segmentation Masks (SegMasks). A few visual

samples for each mode are provided in 3.1. PoseHMs and SegMasks are generated using an off-the-shelf detector [87] without any form of pre/post-processing.

Implementation Details

We choose a 3D-ResNet similar to [20, 24] as the encoder $f(.)$. We choose $N = 8$ and $K = 5$ in our experiments. We subsample the input by uniformly choosing one out of every 3 frames. Our predictive task involves predicting the last three blocks using the first five blocks. We use standard data augmentations during training whose details are provided in later sections. We train our models using Adam [39] optimizer with an initial learning rate of 10^{-3}, decreased upon loss plateauing. We use 4 GPUs with a batch size of 16 samples per GPU. Multiple spatio-temporal samples ensure sufficient negative examples despite the small batch size used for training.

Action Classification

We measure the effectiveness of our learned representations using the downstream task of action classification. We follow the standard evaluation protocol of using self-supervised model weights as initialization for supervised learning. The architecture is then fine-tuned end-to-end using class label supervision. We finally report the fine-tuned accuracies on UCF101 and HMDB51. While fine-tuning, we use the learned composite function $F(.)$ in order to generate context representations for the video blocks. The context feature is further passed through a spatial pooling layer followed by a fully-connected layer and a multi-way softmax for action classification.

3.3.2 Quantitative Results

We analyze various aspects of CoCon through ablation studies, experiments on multiple datasets, controlled variation of modes and comparison to comparable methods. We objectively evaluate model performance using downstream classification accuracy as a proxy for learned representation quality. Pre-training is performed on either UCF101 or Kinetics400. We propose two baselines for comparison. (1) *Random* - random initialization of weights (2) *CPC* - self-supervised training utilizing only $\mathcal{L}_{cpc}$; which is effectively individual training of modes. *CPC* serves as a critical baseline to measure the benefits of multi-modal training as opposed to individual training.

Ablation Study

We have motivated the utility of our various loss components. We now perform experiments to quantify the impact of each. The pre-training dataset used is the 1^{st} split of UCF101, and downstream classification accuracy is computed on the same. Table 3.1 summarizes the results of our experiment. As expected, all cross-view approaches comfortably perform better than *CPC*; demonstrating the utility of multi-modal training.

Method	Resolution	Backbone	# Views	Pre-train	UCF101	HMDB51
Random Initialization	128×128	ResNet18	1		46.7	20.6
ImageNet [75]	224×224	VGG-M-2048	1	ImageNet	73.0	40.5
Shuffle and Learn [56]	227×227	CaffeNet	1	UCF-HMDB	50.2	18.1
OPN [44]	80×80	VGG-M-2048	1	UCF-HMDB	59.8	23.8
DPC [20]	128×128	ResNet18	1	UCF101	60.6	-
VGAN [84]	N/A	C3D	2	Flickr [84]	52.1	-
LT-Motion [52]	N/A	RNN [52]	2	NTU	53.0	-
Cross and Learn [69]	224×224	CaffeNet	2	UCF101	58.7	27.2
Geometry [16]	N/A	CaffeNet	2	UCF101	55.1	23.3
CMC [80]	128×128	CaffeNet	3	UCF101	59.7	26.1
CoCon - RGB	128×128	ResNet18	4	UCF101	**70.5**	**38.4**
CoCon - Ensemble	128×128	ResNet18	4	UCF101	**82.4**	**52.0**
3D-RotNet [35]	112×112	ResNet18	1	Kinetics	62.9	33.7
DPC [20]	128×128	ResNet18	1	Kinetics	68.2	34.5
CoCon - RGB	128×128	ResNet18	2	Kinetics	**71.6**	**46.0**
CoCon - Ensemble	128×128	ResNet18	2	Kinetics	**78.1**	**52.0**
ST-Puzzle [38]	224×224	ResNet18	1	Kinetics	65.8	33.7
DPC [20]	224×224	ResNet34	1	Kinetics	75.7	35.7
CoCon - RGB	224×224	ResNet34	2	Kinetics	**79.1**	**48.5**
CoCon - Ensemble	224×224	ResNet34	2	Kinetics	**82.0**	**53.1**

Table 3.6: Comparison of classification accuracies on UCF101 and HMDB51, averaged over all splits.

Using $\mathcal{L}_{sync}^{cpc}$ leads to no performance improvements, as only using $\mathcal{L}_{sync}$ leads to the model collapsing by squashing all $\mathcal{D}$ scores to have similar values, thus necessitating $\mathcal{L}_{sim}$ to counter-balance this tendency. $\mathcal{L}_{sim}^{cpc}$ leads to improved performance wrt $\mathcal{L}_{cpc}$ as it learns better features by effectively maximizing mutual information between modes. CoCon i.e $\mathcal{L}_{cocon}$ achieves the same by also regularizing manifolds across modes, leading to even better performance across all modes. The important comparison to observe is between $\mathcal{L}_{sim}^{cpc}$ and L_{cocon}. As $\mathcal{L}_{sim}^{cpc}$ is the most similar baseline to other multi-modal approaches, e.g., CMC [80]. However, we argue this baseline is even stronger as it involves both single-view and multi-modal components compared to [80], which only uses a contrastive multi-modal loss to learn representations.

Effect of Datasets

A critical benefit of self-supervised approaches is the ability to run on large unlabelled datasets. To simulate such a setting, we perform pre-training using UCF101 or Kinetics400 [1] without labels utilizing the 1^{st} splits of UCF101 and HMDB51 for evaluation. Table 3.2 confirms pre-training with a larger dataset leads to better performance. It is also worth noting that CoCon pre-trained with UCF101 outperforms *CPC* trained on Kinetics400 even though CoCon on UCF101 uses only around 10% data compared to Kinetics. Further

[1] Optical Flow used for Kinetics400 is Farneback Flow; as opposed to TVL1 Flow for UCF101 and HMDB51. This difference in pre-training and fine-tuning modalities leads to less than expected performance gains.

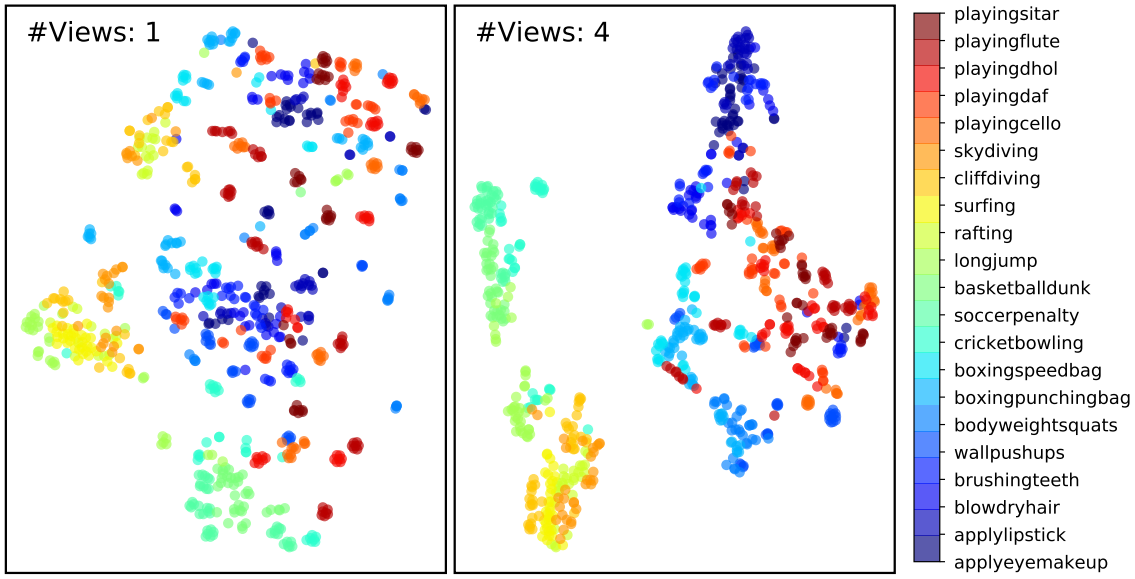


Figure 3.2: t-SNE visualization of RGB features from *CPC* (left) and CoCon (right) trained with 4 modes. The color mapping for each category represents the relationships between action classes, e.g., Red: Instruments; Yellow: Water Sports; Light-blue: Physical Acts; Blue: Makeup-Hygiene. More meaningful clusters are formed using CoCon; signifying the ability of CoCon to align different yet semantically-related classes without any additional supervision.

demonstrating the potential of utilizing multiple modes as opposed to training with larger and diverse datasets.

When comparing the *Random* baseline and CoCon pre-trained on Kinetics400, we observe higher performance gains for RGB (+25.4%) compared to Optical-Flow (+6.9%). We argue this is due to higher variance and complexity of RGB compared to Flow, allowing a randomly initialized network to perform relatively better with Flow. While comparing our approach with *CPC*, we again observe higher gains in RGB (+4.1%) compared to Flow (+2.7%). This can be explained by the potential capability of RGB to capture flow-like features when learned jointly.

Effect of cooperative training

We compare benefits of cooperative training with varying modes. We look at co-training of RGB, Flow, SegMasks and PoseHMs. Recall that these additional modes are generated using off-the-shelf models without any additional post-processing. Even though they are somewhat redundant i.e. Flow, PoseHM, SegMasks are actually derived from RGB Images; using them simultaneously still leads to a large performance increase. We also note that although SegMasks and PoseHMs are sparse low-dimensional features, they still help improve performance across all modes.

Table 3.3 summarizes downstream action recognition performance of each mode under different approaches. We see improved performance with increase in the number of modes used. Consistent gains

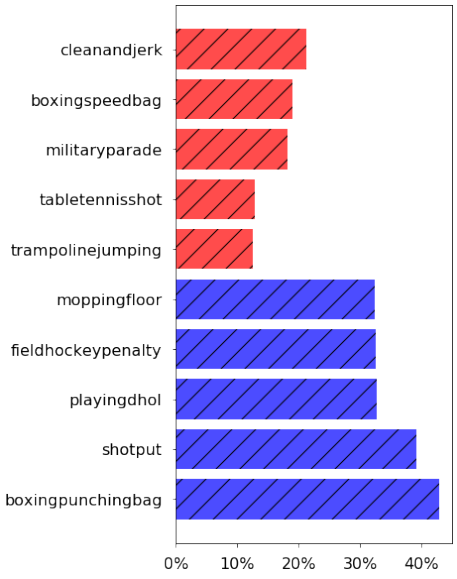


Figure 3.3: Differences between class-wise accuracy for CoCon vs CPC. Only extreme classes are displayed. Blue - Gains; Red - Loss

for modes such as Flow, SegMasks, PoseHM, which are not as expressive as RGB points towards extraction of higher-order features even from low dimensional inputs. We observe PoseHM and SegMask have lower performance gains when evaluated on HMDB51. This can be attributed to the large degree of noise in PoseHMs and SegMasks for HMDB51. HMDB is a challenging and diverse dataset, leading to poor predictions from our off-the-shelf detector. In conclusion, the benefits of joint training are apparent as CoCon leads to a performance improvement for all the modes involved.

Effect of additional modes

CoCon hinges on the assumption that multi-modal information helps in improving overall representation quality. To verify our hypothesis, we study co-training with different number of modes. We consider two scenarios, 1) Joint training of RGB and Flow streams, and 2) Joint training of RGB, Flow, SegMasks and PoseHMs. Table 3.5 shows a consistent increase across modes when increasing the number of modes used during training. We should note that both SegMasks and PoseHMs contain significant noise as the off-the-shelf models incorrectly detects and misses humans in numerous videos. However, we see a consistent mutual increase in performance for all the involved modes despite the prevalence of noise.

Comparison with comparable approaches

We summarize comparisons of CoCon with comparable state-of-the-art approaches in Table 3.6. CoCon-Ensemble refers to an ensemble of models for all the involved modes. We observe a few major trends,

(1) When pre-training on UCF101, using multiple modes allows us to outperform the nearest comparable approach by around 10.4%. This demonstrates the potential of cooperatively utilizing multiple modes to learn representations. (2) We see considerable gains while training on Kinetics400 as well, however, the increase is smaller compared to UCF101. We argue the reasons are, a) we only utilize two modes for co-training. b) the flow we utilize for Kinetics400 is Farneback Flow instead of TVL1 flow used for UCF101 and HMDB51. (3) Our method comfortably outperforms recent multi-modal approaches consistently on UCF101 and HMDB51. (4) An interesting observation is that using multiple modes of a small dataset (UCF101) performs better (71.0%) than pre-training on a large dataset, Kinetics400 (68.2%). This suggests that utilizing different modes can be better than merely training on larger datasets.

Comparison with recent approaches

A few very recent approaches [3, 21, 22, 62] have tackled multi-modal self-supervised achieving impressive performance. CoCon differs from them as it considers inter-instance relationships to aid learning in addition to relationships between modes. Due to resource constraints, it was not possible to have a fair comparison due to the significant difference in the amount of GPUs, number of epochs trained and the backbones used. However, we hope our carefully constructed experiments given earlier provide deeper insights into CoCon's benefits even with lower resource requirements.

Figure 3.4: Soft Alignment of videos from UCF101 test split using CoCon pre-trained on UCF101. The first pair of videos involves pull-ups; observe the periodicity captured in the heatmap. The second involves high-jumps; notice that we are roughly able to align the running and jumping phases though they happen at different times. Heatmaps (right) represent relative block similarities from different time-steps of the videos. The color of the frame boxes describe the associated actions; matching colors broadly represent similar action stages.

3.3.3 Qualitative Results

We motivate CoCon arguing about the benefits of preserving similarities across mode-specific spaces. We observe respecting structure across modes results in emergence of higher-order semantics without additional supervision e.g. sensible class relationships and good feature representations. Jointly training with modes known to perform well for video action understanding allows us to learn good video representations, consequently, imparting unexpected side effects such as action alignment across videos. We discuss various experiments and results to support these claims.

t-SNE Visualization

We explore t-SNE visualizations of our learned representations on the 1^{st} test split of UCF101 extracted using $F(.)$. For clarity, only 21 action classes are displayed. We loosely order the action classes according to their relationships. Classes having similar colors are semantically similar. We can roughly observe the following broad categories present in the mentioned classes: Playing Instruments, Water Sports, Physical Sports, Physical Activities, Makeup-Hygiene. Results are displayed in Fig 5.3. Although we operate in a self-supervised setting, CoCon is able to uncover deeper semantic features allowing us to uncover inter-class relationships. We can see a much more concise and consistent clustering in CoCon compared to CPC.

Effect of action classes on performance

Figure 3.3 shows the classes which observe the least and highest performance improvements when co-trained with multiple modes. We observe a loose pattern where action classes involving distinguishable physical movements see larger improvements. We can argue this is because we use modes which are suitable for physically intensive actions.

Inter-Class Relationships

In order to study consistency of structure across different modes, we look at relationships between classes by inferring their similarities through our learned features. We compare cosine similarities across videos from different classes and compute the most similar four classes for each action. We repeat the process for all modes and look at the consistency of the results. We only display classes which are amongst the closest ones across all modes. Table 3.4 summarizes our results. We see the detected nearest actions are semantically related to the original actions. In the cases of PlayingCello, we encounter a cluster of categories involving playing instruments. Similarly for BasketBall, we can see emergence of sports-based relationships even though there is not any visual commonality between the categories. It's worth noting that as these nearest classes are consistent across different modes, our approach cannot cheat to generate them i.e. it cannot look at 'background crowd' or 'green field' and infer that a video clip is related to sports. Since modes such as Optical-Flow, SegMasks and KeypointHeatmap do not have such information and are very low-dimensional.

Action Alignment

Even though we only use self-supervision, our embeddings are able to capture relevant semantics through our multi-modal approach allowing loose alignment between videos. To compute this soft alignment, we divide each video into 18 blocks and compute block-level features. We then utilize relative cosine similarities to infer associations between the videos. Figure 3.4 highlights a few examples. Notice the periodicity implicitly present in some actions (e.g. pullups) captured through the heatmap allowing us to perform non-linear alignment.

3.4 Conclusion

We propose a cooperative version of contrastive learning, called CoCon, for self-supervised video representation learning. By leveraging relationships across modes, we encourage our self-supervised learning objective to be aligned with the underlying semantics. We demonstrate the effectiveness of our approach on the downstream task of action classification, and illustrate the semantic structure of our representation. We show that additional input modes generated by off-the-shelf computer vision algorithms can lead to significant improvements, even though they are noisy and derived from an existing modality i.e. RGB. As these modes are 'freely' available, this shows the feasibility of utilizing multi-modal approaches on datasets which are not traditionally considered multi-modal. We hope this enables the ability to leverage multi-modal learning algorithms and observe performance gains even on single-view datasets.

3.5 Additional Details

3.5.1 Model Overview

We build our framework borrowing the learning framework present in [20] which learns video representations through spatio-temporal contrastive losses. It should be noted that even though we use this particular self-supervised backbone in our experiments, our approach is not restricted by the choice of the underlying self-supervised task.

A video V is a sequence of T frames (not necessarily RGB images) with resolution $H \times W$ and C channels, $\{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_T\}$, where $\mathbf{i}_t \in \mathbb{R}^{H \times W \times C}$. Assume $T = N * K$, where N is the number of blocks and K denotes the number of frames per block. We partition a video clip V into N disjoint blocks $V = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_j \in \mathbb{R}^{K \times H \times W \times C}$ and a non-linear encoder $f(.)$ transforms each input block x_j into its latent representation $z_j = f(x_j)$.

An aggregation function, $g(.)$ takes a sequence $\{z_1, z_2, \ldots, z_j\}$ as input and generates a context representation $c_j = g(z_1, z_2, \ldots, z_j)$. In our setup, $z_j \in \mathbb{R}^{H' \times W' \times D}$ and $c_j \in \mathbb{R}^D$. D represents the embedding size and H', W' represent down-sampled resolutions as different regions in z_j represent features for different

spatial locations. We define $z'_j = Pool(z_j)$ where $z'_j \in \mathbb{R}^D$ and $c = F(V)$ where $F(.) = g(f(.))$. In our experiments, $H' = 4, W' = 4, D = 256$.

To learn effective representations, we create a prediction task involving predicting z of future blocks similar to [20]. In the ideal scenario, the task should force our model to capture all the necessary contextual semantics in c_t and all frame level semantics in z_t. We define $\phi(.)$ which takes as input c_t and predicts the latent state of the future frames. The formulation is given in Eq. (5.3).

$$\begin{aligned} \widetilde{z}_{t+1} &= \phi(c_t), \\ \widetilde{z}_{t+1} &= \phi(g(z_1, z_2, \ldots, z_t)), \\ \widetilde{z}_{t+2} &= \phi(g(z_1, z_2, \ldots, z_t, \widetilde{z}_{t+1})), \end{aligned} \tag{3.4}$$

where $\phi(.)$ takes c_t as input and predicts the latent state of the future frames. We then utilize the predicted $\widetilde{z}_{t+1}$ to compute $\widetilde{c}_{t+1}$. We can repeat this for as many steps as we want, in our experiments we restrict ourselves to predict till 3 steps in to the future.

Note that we use the predicted $\widetilde{z}_{t+1}$ while predicting $\widetilde{z}_{t+2}$ to force the model to capture long range semantics. We can repeat this for a varying number of steps, although the difficulty increases tremendously as the number of steps increases as seen in [20]. In our experiments, we predict the next three blocks using the first five blocks.

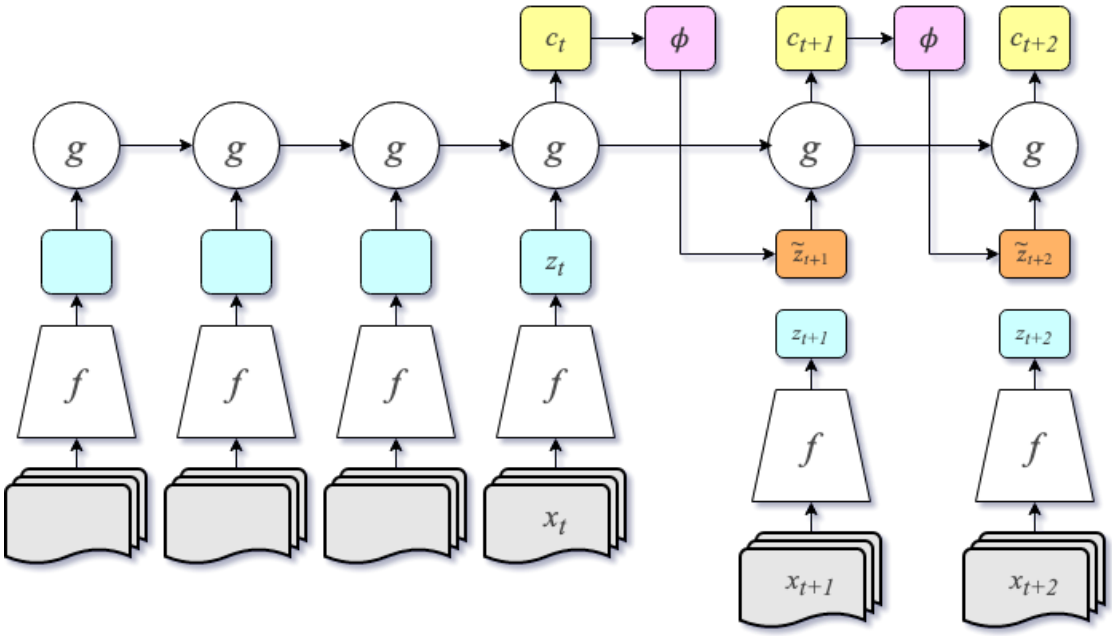


Figure 3.5: A diagram of the learning framework utilized. We look at features in a sequential manner while simultaneously trying to predict representations for future states.

3.5.2 Datasets

Kinetics400 contains 400 human action classes, with at least 400 real-world video clips for each action. Each clip lasts around 10s and is taken from a different YouTube video. The actions are human focused and cover a broad range of classes including human-object and human-human interactions. The large diversity and variance in the dataset make it an extremely challenging dataset.

HMDB51 dataset contains around 6800 real-world video clips from 51 action classes. These action classes cover a wide range of actions - facial actions, facial action with object manipulations, general body movement, and general body movements with human interactions. This dataset is challenging as it contains many poor quality video with significant camera motions and also the number of samples are not enough to effectively train a deep network. We report classification accuracy for 51 classes across 3 splits provided by the authors.

UCF101 dataset contains 13320 videos from 101 action classes that are divided into 5 categories - human-object interaction, body-movement only, human-human interaction, playing musical instruments and sports. Action classification in this datasets is challenging owing to variations in pose, camera motion, viewpoint and spatio-temporal extent of an action.

3.5.3 Views

We simultaneously learn encoders for RGB and Optical Flow while training on Kinetics-400. Instead of using the commonly used TVL1-Flow, we rely on Farneback flow which usually results in lower performance for action classification, however is much faster to compute. For UCF101 and HMDB51, we simultaneously learn encoders for RGB, TVL1 Optical Flow, Pose Heatmaps and Semantic Maps.

We give a brief overview of the views utilized and their generation.

- **RGB Images, RGB** - We directly use sequences of RGB frames present in videos
- **Optical Flow, Flow** - We use the popular TVL1 flow for UCF101 and HMDB51 and Farneback Flow (FF) for Kinetics400. FF is known to perform worse than TVL1-Flow on visual recognition tasks, however, it is quicker to compute. This view mismatch views leads to lesser gains when using Kinetics pre-trained flow weights for UCF101 and HMDB51.
- **Pose Keypoint Heatmaps, PoseHMs** - We use an off-the-shelf keypoint detector [87] and extract confidence heatmaps for each keypoint. Note that we perform no pre/post-processing on the results and directly use this as input to our model. The input modality is inherently very noisy, however, we still observe improved performance.
- **Human Segmentation Masks, SegMasks** - Similar to the above, we use an off-the-shelf semantic segmentation network [87] and extract confidence scores for human segmentation. Similar to pose keypoint heatmaps, this input modality is inherently very noisy.

Fig. 3.6 shows examples of different views. Note the prevalence of noise in a few samples, specially SegMasks. There are multiple other instances where PoseHMs are noisy as we're unable to even localize the actor accurately.

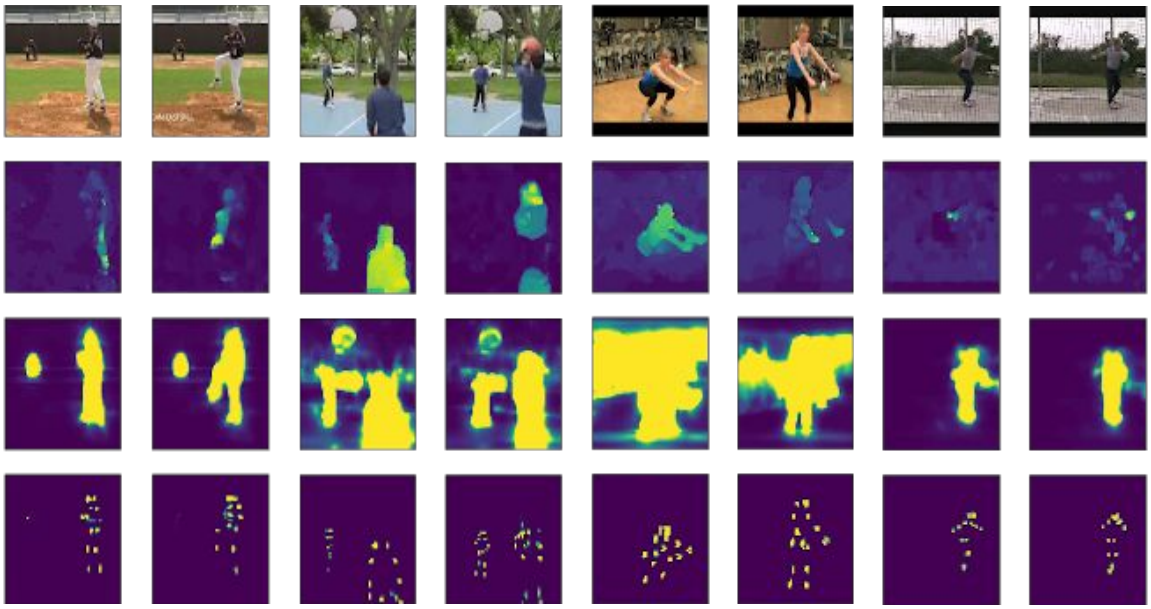


Figure 3.6: Examples for each view. From top to bottom - RGB, Flow, SegMasks and Poses.

3.5.4 Implementation Details

We choose to use a 3D-ResNet similar to [24] as the encoder $f(.)$. Following [20] we only expand the convolutional kernels present in the last two residual blocks to be 3D ones. We used 3D-ResNet18 for our experiments, denoted as ResNet18. We use a weak aggregation function $g(.)$ in order to learn a strong encoder $f(.)$. Specifically, we use a one-layer Convolutional Gated Recurrent Unit (ConvGRU) with kernel size (1, 1) as $g(.)$. The weights are shared amongst all spatial positions in the feature map. This design allows the aggregation function to propagate features in the temporal axis. A dropout [27] with $p = 0.1$ is used when computing the hidden state at each time step. A shallow two-layer perceptron is used as the predictive function $\phi(.)$. Recall $z'_j = Pool(z_j) where z'_j \in R_D$. We utilize stacked max pool layers as $Pool(.)$.

To construct blocks to pass to the network, we uniformly choose one out of every 3 frames. We then group these into 8 blocks containing 5 frames each. Since the videos we use are usually 30fps, each block roughly covers 0.5 seconds worth of content. The predictive task we design involves predicting the last three blocks using the first five. Therefore, we effectively predict the next 1.5 seconds based on the first 2.5 seconds.

We perform random cropping, random horizontal flipping, random greying, and color jittering to perform data augmentation in the case of images. For optical flow, we only perform random cropping on the image. As

discussed earlier, Keypoint Heatmaps and Segmentation Confidence Masks are modelled as images, therefore we perform random cropping and horizontal flipping in their case. Note that random cropping and flipping is applied for the entire block in a consistent way. Random greying and color jittering are applied in a frame-wise manner to prevent the network from learning low-level features such as optical flow. Therefore, each video block may contain both colored and grey-scale image with different contrast.

All individual view-specific models are trained independently using only $\mathcal{L}_{cpc}$. After which we proceed to train all view-specific models simultaneously using $\mathcal{L}_{cocon}$. All models are trained end-to-end using Adam [39] optimizer with an initial learning rate 10^{-3} and weight decay 10^{-5}. Learning rate is decayed to 10^{-4} when validation loss plateaus. A batch size of 16 samples per GPU is used, and our experiments use 4 GPUs. We train models on UCF101 for 100 epochs using $\mathcal{L}_{cpc}$, after which they are collectively trained together for 60 epochs using $\mathcal{L}_{cocon}$. We repeat the same for Kinetics400 with reduced epochs. We train models on Kinetics400 for 80 epochs using $\mathcal{L}_{cpc}$ and further for 40 epochs using $\mathcal{L}_{cocon}$.

The learned representations are evaluated by their performance on the downstream task of action classification. We follow the evaluation practice from recent works and use the weights learned through our self-supervised framework as initialization for supervised learning. The whole setup is then fine-tuned end-to-end using class label supervision. We finally report the fine-tuned accuracies on UCF101 and HMDB51. We use the learned composite function $F(.)$ to generate context representations for video blocks. The context feature is further passed through a spatial pooling layer followed by a fully-connected layer and a multi-way softmax for action classification. We use dropout with $p = 0.7$ for classification. The models are fine-tuned for 100 epochs with learning rate decreasing at different steps. During inference, video clips from the validation set are densely sampled from an input video and cut into blocks with half-length overlapping. The softmax probabilities are averaged to give the final classification result.

Additional Results

We motivate CoCon arguing about the benefits of preserving similarities across view-specific feature spaces. We observe respecting structure across views results in emergence of higher-order semantics without additional supervision e.g. sensible class relationships and good feature representations. We go over different results in the following sections.

3.5.5 t-SNE Visualization

We explore t-SNE visualizations of our learned representations on the 1^{st} test split of UCF101 extracted using $F(.)$. Our model is trained on the corresponding train split to ensure we're testing out of sample quality. For clarity, only 21 action classes are displayed. We loosely order the action classes according to their relationships. Classes having similar colors are semantically similar. Results are displayed in Fig 3.7. Even though we operate in a self-supervised setting, our approach is able to uncover deeper semantic features allowing us to uncover inter-class relationships. We can see a much more concise and consistent clustering

in CoCon compared to CPC. We also observe the distinct improvement in the compactness of the clusters as we increase the number of views.

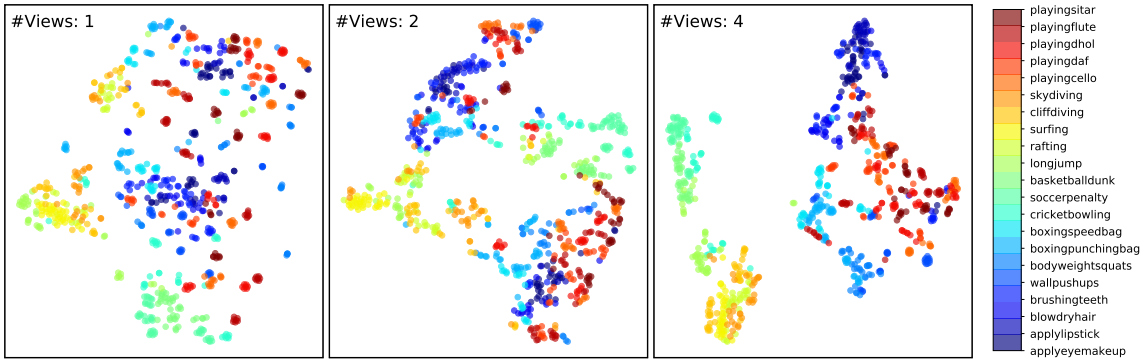


Figure 3.7: Emergence of relationships between different actions using CoCon with varying number of views. Note that CoCon becomes the same as CPC when $\#views = 1$

3.5.6 Inter-Class Relationships

In order to study the manifold consistency across different views, we look at relationships between classes by inferring their similarities through the learned features. We compare cosine similarities across video clips from different classes. We then compute the most similar five classes for each action. We repeat the process for all views and look at the consistency of the results. Ideally, semantically similar classes should be consistent across all views, assuming the views reasonably capture the essence of the task we're interested in.

We observe that CoCon leads to much higher consistency across different views. Specifically, we see 41 classes which have at least four out of five top-classes consistent in all views; as opposed to 10 classes in CPC. Similar patterns are seen when we consider other thresholds. In order to confirm that the nearest classes are actually sensible, we mention the most-similar classes for a few action classes.

We can see that the nearest actions generated are semantically related to the original actions. In the cases of PlayingCello, we encounter a cluster of categories involving playing instruments. Similarly for BasketBall, we can see emergence of sports-based relationships even though there is no visual commonality between categories. We also see a few seemingly unrelated classes as well, e.g., BoxingPunchingBag and YoYo; SalsaSpin and WalkingWithDog. A deeper inspection into the samples is required to comment whether this truly makes sense. It is worth noting that as these nearest action classes are mostly consistent across different views, our approach cannot cheat to generate them i.e. it cannot look at 'background crowd' or 'green field' and infer that the video clip is related to sports. Since views such as Optical-Flow, SegMasks and KeypointHeatmap do not have such information and are much low-dimensional.

Action Class	Nearest Classes	
	CoCon	CPC
skiing	surfing, skijet	surfing
playingcello	playingsitar, playingtabla, playingdhol	N/A
jumpingjack	jumprope, pullups, bodyweightsquats, cleanandjerk	N/A
basketball	baseballpitch, cricketshot, fieldhockey, cricketbowling	N/A
hammerthrow	baseballpitch, throwdiscus, shotput	N/A
wallpushups	writingonboard, bodyweightsquats	N/A
brushingteeth	applylipstick, applyeyemakeup, shavingbeard, haircut	applylipstick

Table 3.7: Closest semantic classes provided by different models. CPC has very few consistent nearest classes across views. While views trained using CoCon show consistent results across views, leading to sensible inter-class relationships

3.5.7 Action Alignment

An interesting side-effect of improved representations for actions is the possibility of performing loose action alignment. Even though we only use self-supervision, CoCon embeddings are able to capture relevant semantics through our multi-view approach allowing loose alignment between videos. To compute this soft alignment, we divide each video into 18 blocks and compute block-level features z'. We then utilize relative cosine similarities to infer associations between the videos. We smoothen the heatmap in order to make it visually appealing. Figure 3.4 shows alignment between different videos. Figure 3.8 highlights a few examples when we perform alignment between same videos. Notice the periodicity implicitly present in these actions captured through the heatmap.

3.5.8 Cosine similarity

This section highlights the ability of representation generated through CoCon to capture meaningful semantics going beyond low-level features. We look at cosine similarity distributions of video representation from UCF101. We extract one context representation for each video and pool it into a vector. We then compute the cosine similarity for each pair of video features across the unseen UCF101 test set. The cosine distance is summarized by a histogram, where the 'blue' histogram represents the score distribution for positives i.e. videos belonging to the same class; and the 'orange' one shows the distribution for negatives i.e. videos from different classes.

3.5.9 Nearest Neighbors

We utilize CoCon to perform video retrieval for different query videos. Note that CoCon is able to look past purely visual background features and focus on actions even though it only used RGB inputs. For example, we see that we are able to retrieve close neighbors for BenchPress, even though it is very visually different with varying poses. For the IceDancing sample, even though it incorrectly considers onbe video where the person is running, we can still see similarities between the underlying actions in the videos. Similar results

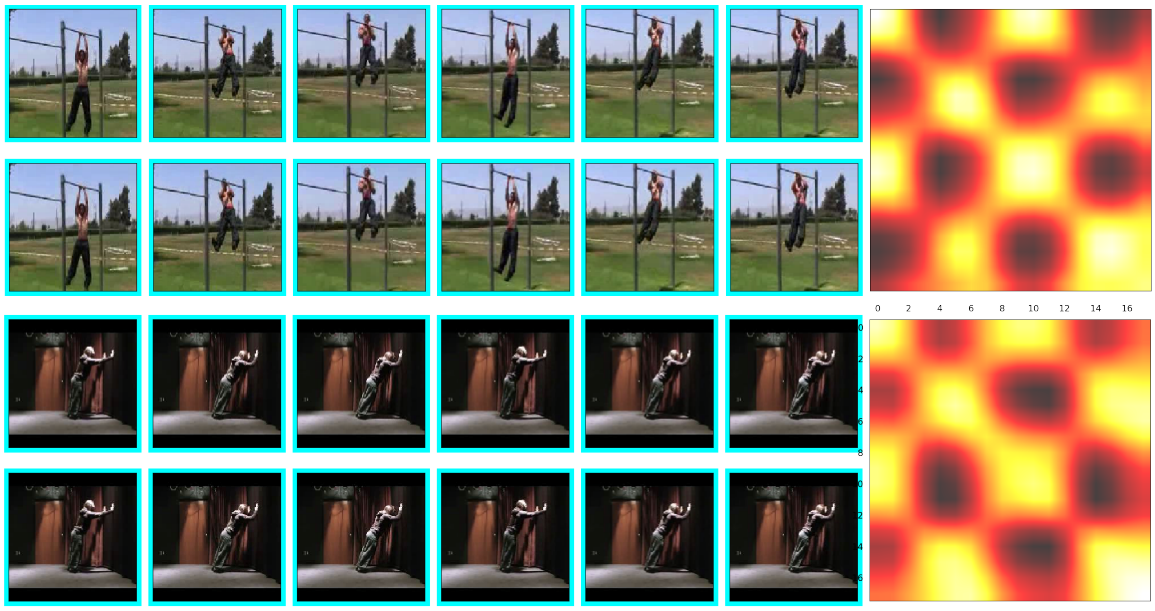
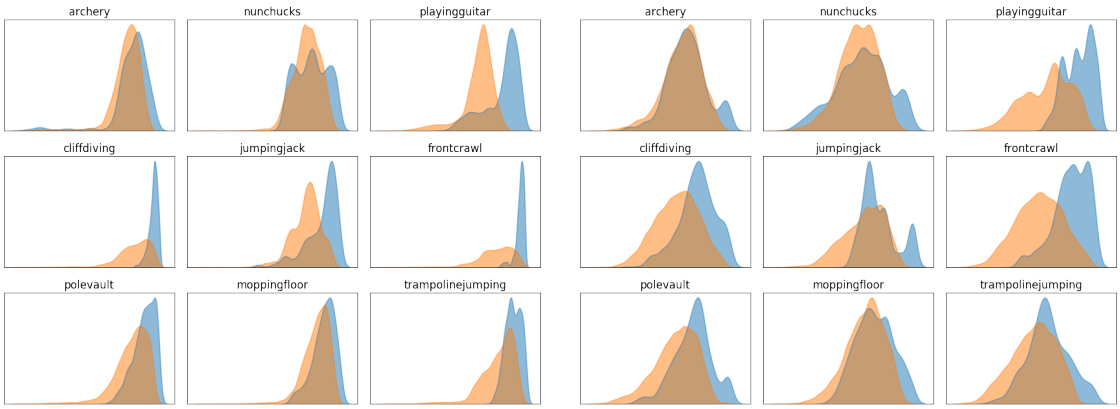


Figure 3.8: Soft Alignment of actions between the same video instances. The heat-map represents the relative similarities between blocks at various timesteps. Notice periodic patterns in the actions.

can be seen in other examples as well. This hints towards the fact that CoCon representation are able to capture action semantics even while using RGB views.



(a) Generated by CoCon (b) Generated by CPC

Figure 3.9: Distributions of cosine-similarity scores between representations of videos from the same (blue) and other classes (red).

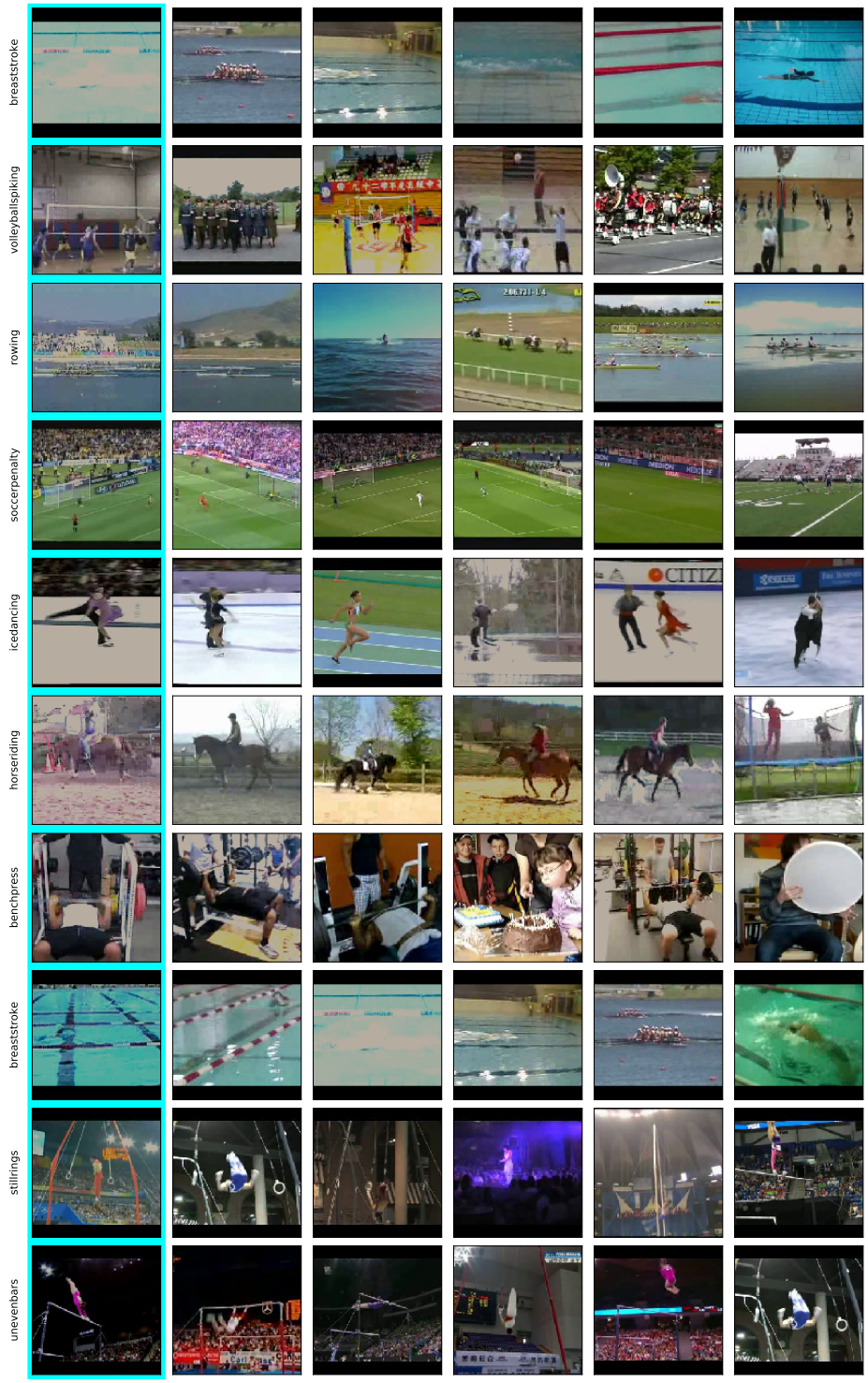


Figure 3.10: Nearest neighbors computed using RGB representations. Query video is highlighted on the left with Aqua Blue.

Chapter 4

Home Action Genome

4.1 Introduction

Action understanding in videos is a critical task with various use-cases and real-world applications, from robotics [47, 66] and human-computer interaction [68] to healthcare [23, 50] and elderly behavior monitoring [30, 51]. Despite the recent success of deep learning methods for image classification, complex and holistic action or event understanding remains an elusive task.

There are several challenges associated with the task of action understanding. The inherent variability in executing complex activities poses one of the most critical difficulties in building action understating models. To understand these challenges, it is essential to understand what actions are composed of. As opposed to bounding boxes in the object detection task, actions are composed of various parts spanned in space and time. For instance, the action of "laundry" involves multiple entities, e.g., humans, objects, and their relationships, and is composed of a number of atomic actions. Such partonomy of actions [5, 32, 89] both in space and time defines a hierarchical structure. Furthermore, to capture the variability in executing complex activities, understanding each part (e.g., body limbs, objects, or atomic actions) becomes crucial. Since actions happen in the 3D world, a holistic understanding of the world requires capturing the subtle movements or parts using multiple modalities (e.g., RGB and audio) and from multiple viewpoints.

Each of these challenges has previously been separately investigated using different datasets and advanced methods. For instance, numerous datasets were put together for generic action recognition and spatio-temporal localization in YouTube or broadcasting third-person videos, such as Kinetics [6], Charades [72], ActivityNet [14], UCF101 [76]. Other datasets such as EPIC Kitchens [10] were used for ego-centric action recognition. Action Genome [32] focused on using scene information in action recognition, while others [53] focused on hierarchical action modeling from events to low-level atomic actions. Several studies target learning from long instructional videos and release datasets [8, 55, 79, 91] for the same, exploring the partonomy of actions in long sequences. Others also focused on observing and recognizing actions from multiple views, such as LEMMA [34] and HumanEva [71]. In parallel, there have been numerous recent

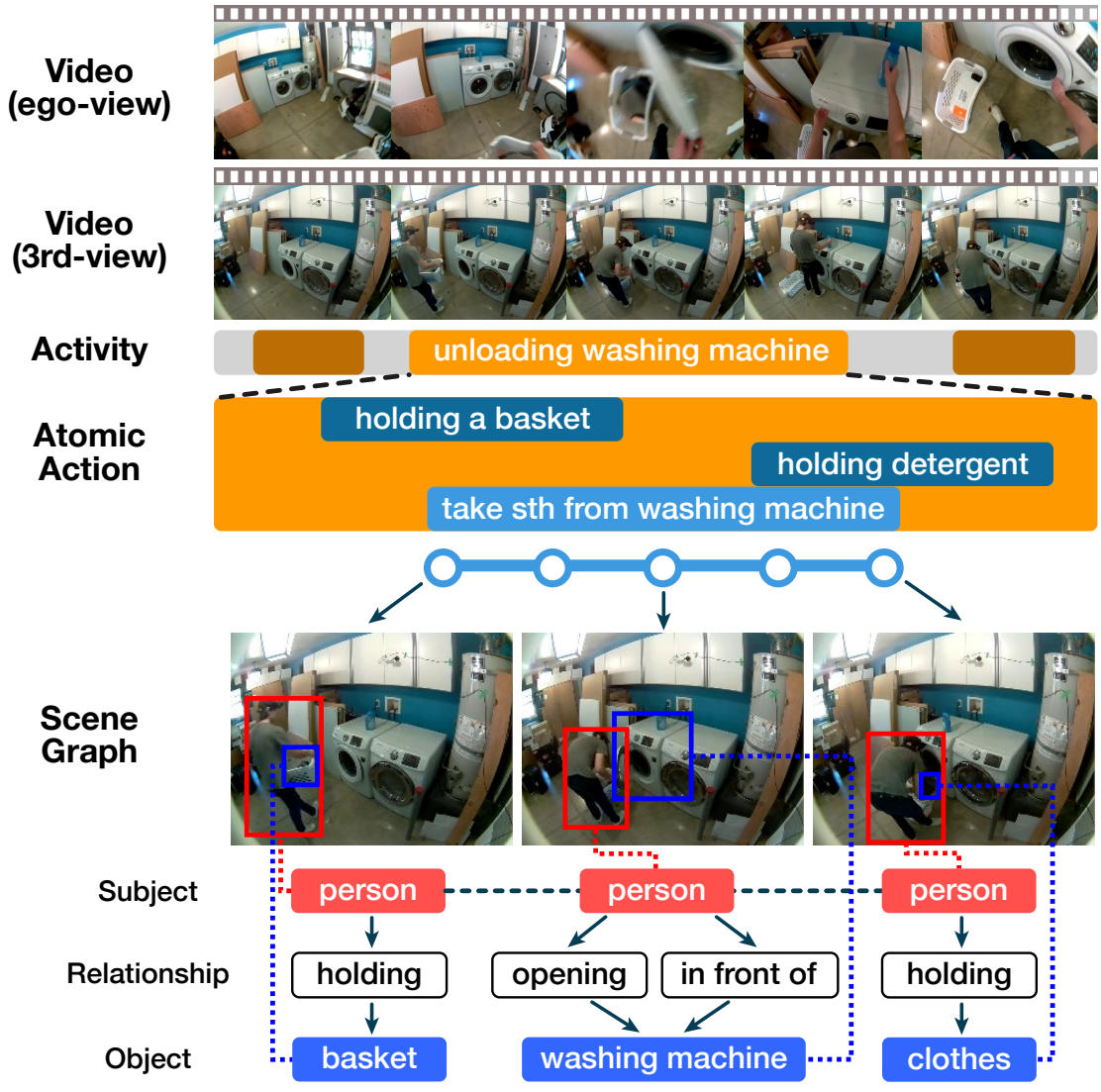


Figure 4.1: HOMAGE annotation pipeline: For every action, we uniformly sample 3 or 5 frames across the action and annotate the bounding boxes of the person performing the action along with the objects they interact with. We also annotate the pairwise relationships between the subject and the objects.

advances in contrastive and cooperative learning [9, 22] applied to multi-modal and multi-view datasets as a self-supervised pre-training strategy to improve downstream recognition results. Despite all these advances, action understanding and generalizability of such models remains a challenging problem due to complexities brought by their complicated nature and numerous object interactions. Multi-modal approaches [69, 74, 80] have shown superior performance in tackling such issues. However, there is still a need for a benchmark that unifies all these challenges and tasks. In this paper, we release a dataset along with a novel method for hierarchical action recognition to tackle these problems.

We introduce a new benchmark for action recognition, Home Action Genome (HOMAGE), that includes multi-modal synchronized videos from multiple viewpoints along with hierarchical action and atomic-action labels. Actions in homes are challenging as we deal with long-term actions, interactions with objects, and frequent occlusions. Having multiple views and sensors to handle occlusions and scene graph information to capture object interaction allows us to tackle these complexities. In addition, synchronous videos provide implicit alignment that facilitates multi-modal training. Additionally, access to sensor information enables future research in privacy-aware recognition where we avoid audio-visual modalities. HOMAGE also provides temporal annotations of high-level activity and low-level atomic action supplemented with spatio-temporal scene-graphs. Annotations regarding interaction of objects within actions and atomic actions within high-level actions enable research in explainable video understanding, early action prediction, and long-range action recognition.

As we will see in a later chapter, for this new benchmark, we introduce a novel method to perform simultaneous co-training with multiple modalities (RGB, audio, and annotations of scene composition) and viewpoints that enable the learning of rich video representations. Training involves a co-training strategy that leverages information from all views and modalities to build the representation space. During inference, we set up different experiments and observe improved action recognition performance even when only a single modality is used, which suggests training on HOMAGE improves performance with no need for other modalities during inference. In this paper, we explore audio-visual data (of interest to the vision community). Future sensor-fusion work can further exploit other modalities we release (e.g., for privacy-preserving studies).

HOMAGE aims to unify various aspects and challenges of action recognition, specifically targeting multi-modal and compositional perception for home actions. Moreover, the presence of a large number of modalities in our dataset encourages research in areas such as privacy-aware recognition and sensor-fusion. To summarize, our contributions are as follows:

(1) We introduce a new dataset, Home Action Genome (HOMAGE) with multiple views and modalities densely annotated with scene graphs and hierarchical activity labels (overall activity and atomic actions).

(2) We propose a novel learning framework (CCAU) that leverages multiple modalities and hierarchical action labels and improves the performance of the baselines trained on each individual modality. We demonstrate the benefits of our approach with an improvement of +6.4% using only ego-view during inference.

This chapter discusses the first contribution regarding the dataset, how it is collected, its specifics and its

potential applications. We focus on the second contribution in the next chapter.

4.2 Related Work

4.2.1 Action Recognition in Videos.

Action recognition has continuously been an important direction for the computer vision research community. The success of 2D convolutions in image classification allowed frame-level action recognition to become a viable approach. Subsequently, two-stream networks for action recognition [74] have led to many competitive approaches, which demonstrates using multiple modalities such as optical flow helps improve performance considerably. Their work motivated other approaches that model temporal motion features together with spatial image features from videos. [81, 82] demonstrated that replacing 2D convolutions with 3D convolutions leads to further performance improvements. Recent approaches such as I3D [7] inflate a 2D convolutional network into 3D to benefit from the use of pre-trained models. 3D-ResNet [24] adds residual connections building a very deep 3D network leading to improved performance.

4.2.2 Related Datasets

MSR-Action3D [45] provides depth map sequences containing 20 actions of interactions with game consoles. [48, 58, 70, 78] use the Microsoft Kinect sensor to collect multi-modal action data with RGB and depth map sequences. NTU RGB+D [70] consists of RGB, depth map, infrared frames with 3D human joints annotations with 40 human subjects, and 80 distinct camera viewpoints. However, for action labels, each video in these datasets has a single video-level label and thus tough to use for action localization applications.

Other datasets [14, 32, 34, 40, 48] provide annotations for temporally localized actions. MMAct [40] is a large-scale action recognition benchmark multimodal data including RGB videos, keypoints, acceleration, gyroscope, and orientation. It provides an ego-view and 4 third-person views and temporally localized actions. However, MMAct does not provide bounding box annotations for spatial localization and relationships between objects. LEMMA [34] is a recent multi-view and multi-agent human activity recognition dataset, providing bounding box annotations on third-person views and compositional action labels annotated with predefined action templates and verbs/nouns. However, they do not provide bounding boxes of objects the subjects (human) interact with. Action Genome [32] is built upon the videos from Charades [73], with the additional annotation of spatio-temporal scene graph labels. However, it only provides videos from a single camera view. HOMAGE aims to provide 1) multiple modalities to promote multi-modal video representation learning, 2) high-level activity labels and temporally localized atomic action labels, and 3) scene graphs that provide spatial localization cues for both the subject and the object and their relationship.

4.2.3 Multi-Modal Learning.

Multiple modalities of videos are rich sources of information for both supervised [74] and self-supervised learning [69, 80, 84]. [41, 80] introduce a contrastive learning framework to maximize the mutual information between modalities in a self-supervised manner. The method achieves state-of-the-art results on unsupervised learning benchmarks while being modality-agnostic and scalable to any number of modalities. Two stream networks for action recognition [74] have led to many competitive approaches, which demonstrate using even derivable modalities such as optical flow helps improve performance considerably. There have been approaches [52, 69, 80, 84] utilizing diverse modalities, sometimes derivable from one other, to learn better representations.

4.3 Home Action Genome (HOMAGE)

Home Action Genome (HOMAGE) is a new benchmark for action recognition that includes multi-modal synchronized video data from multiple viewpoints (ego-view, third-person) with both high-level activity and low-level action definitions. HOMAGE focuses on actions in residential settings due to the challenges involved i.e. complexity and long duration of actions, object interactions, and frequent occlusions. HOMAGE provides multiple views and sensors to tackle these challenges. We describe the design, data collection, and data annotation process of the HOMAGE dataset in this section.

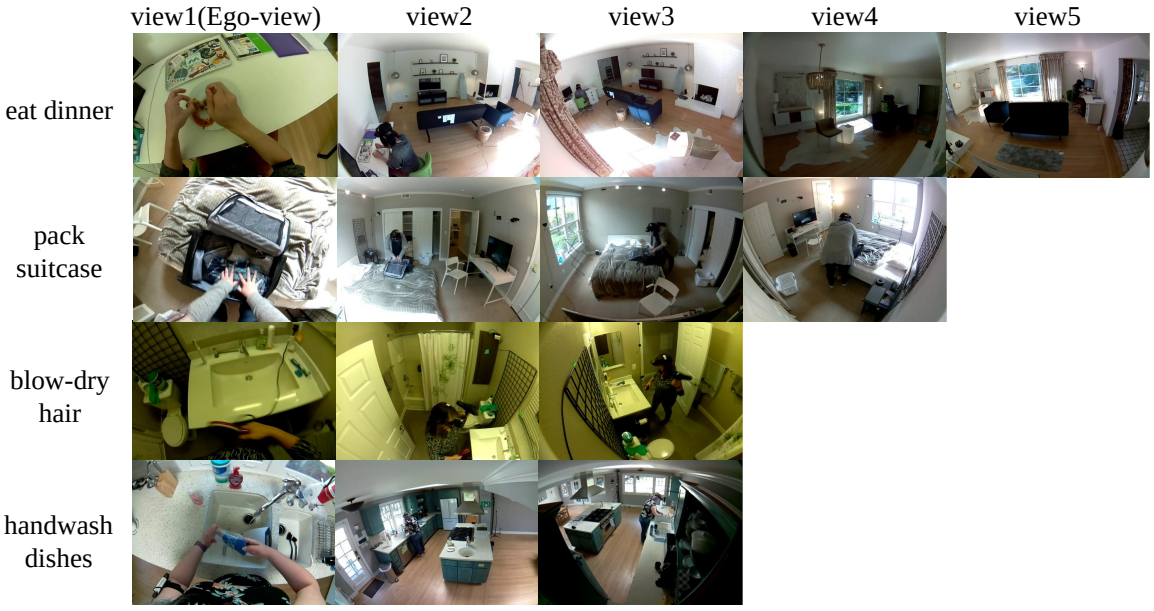


Figure 4.2: Multiple Views of Home Action Genome (HOMAGE) Dataset. Each sequence has one ego-view video as well as at least one or more synchronized third person views.

Dataset	Seq	hrs	Modalities	Views	HL	HL Classes	TL	TL Classes	TL Ins	SG
RGBD-HuDaAct [58]	1.19K	46	2	1	✓	12	-	-	-	-
UCF101 [76]	13K	27	1	1	✓	101	-	-	-	-
ActivityNet [14]	28K	648	1	1	✓	200	-	-	-	-
Kinetics-700 [6]	650K	1.79K	1	1	✓	700	-	-	-	-
AVA [17]	430	108	1	1	-	-	✓	80	1.58M	-
PKU-MMD [48]	1.08K	50	3	3	-	-	✓	51	20K	-
EPIC-Kitchens [10]	-	55	1	1	-	-	✓	125	39.6K	-
MMAct [40]	36K	-	6	5	-	-	✓	37	36.8K	-
Action Genome [32]	10K	82	1	1	-	-	✓	157	66.5K	✓
Breakfast [42]	-	77	1	1	✓	10	✓	48	-	-
LEMMA [34]	324	10.1	2	4	✓	15[1]	✓	863	11.8K	-
Ours	1.75K	25.4	12	2∼5	✓	75	✓	453	24.6K	✓

Table 4.1: Comparison between related datasets and HOMAGE. (Seq: number of synchronized sequences, Modalities: sensor modalities not including annotation data or derived data like optical flow, Views: number of synchronized viewpoints for a given sample, HL: high-level activity label (often assigned one per video), TL: temporally localized atomic action label, SG: scene graph). HOMAGE provides rich multi-modal action data, including dense annotations such as scene graphs, along with hierarchical action labels.

Sensor	Sensor information	
	Model no.	rate
Video	OmniVision OV5647	30fps
I2S Digital Microphones	SPH0645LM4H	48KHz
GridEYE Thermal Imager	AMG8833	10Hz
Human Presence (PIR)	AK9753AE	2Hz
Ambient Light Intensity	TSL2591	2Hz
Ambient Color	ISL29125	10Hz
CO2/Humidity/Pressure/Temp.	BME680	5Hz
Magnetometer	MLX90393	10 Hz

Table 4.2: List of sensors in our multi-modal sensor

4.3.1 Activities and Scenarios.

Our goal is to build an activity recognition dataset that depicts behaviors observed in living spaces. To cover daily activities, we employed the activity taxonomy in the American Time Use Survey (ATUS) [19]. The ATUS taxonomy organizes activities according to two key dimensions: 1) social interactions and 2) the locations of the activities. The ATUS coding lexicon contains a large variety of daily human activities organized under 18 top-level categories such as Personal Care, Work-Related, Education, and Household activities.

Each participant was asked to perform tasks according to the instructions assigned. To make sure the behaviors are as natural as possible, we did not specify detailed procedures and time limits within the activities, and let the individual participants perform the activity freely.

4.3.2 Data Collection.

We recorded 27 participants in kitchens, bathrooms, bedrooms, living rooms, and laundry rooms in two different houses. We used 12 sensor types: cameras (RGB), infrared (IR), microphone, RGB light, light, acceleration, gyro, human presence, magnet, air pressure, humidity, temperature. We refer to the set of data collected from a given activity with different modalities as one synchronized action sequence. Sensors were attached to several locations in the room for third-person views and to the participants' heads for ego-view. On average, there are more than 3 views per action sequence. We synchronized the sensor recordings of all views giving us synced videos which allowed for ease of use without requiring any additional post-processing.

We collect human action data from different viewpoints using our multimodal sensor kits. We provide additional details in Table **??**. Specifically, we synchronize the data from different modalities by using the following scheme. (1) The participants were instructed to start the activity displayed on the screen after they heard the start tone. (2) The content of the participants was specified by activity unit (e.g. make bed). We do not specify a detailed sequence of atomic actions. (3) We sounded the end tone when the participant's activity is finished. We synchronized the data of multiple sensor-kits using the signal of start/end tone.

It is worth noting that in order to measure natural activities in a situation where we are in control of the collecting location and the objects, we did not give instructions to the actors as much as possible. We do not provide any sequence of actions or objects to touch. Furthermore, to match the activity labels like "make bed", the activity instructions were presented in text by display, and the actors did what they could imagine with the activity.

4.3.3 Ground-truth Annotation.

Home Action Genome is a dataset with (1) video-level activity labels, (2) temporally localized atomic activity labels, and (3) spatio-temporal scene-graph labels. Figure 4.1 visualizes our annotation pipeline. For the atomic actions, we annotated all atomic actions performed during the activities. Note that while each video can only have a single activity label, a given frame can be assigned with multiple atomic action labels when atomic actions overlap with each other. For the action graph, we annotated the person performing the action and the objects they interact with on videos from third-person views. We uniformly sampled 3 or 5 to annotate scene graphs across the range of each atomic action interval (3 for intervals less than 3 seconds and 5 otherwise). This action-oriented dynamic sampling provides more labels where more actions occur which is very valuable for describing complex primitive actions. [33] also shows this sampling scheme performs remarkably well.

4.3.4 Dataset Statistics.

We annotated 75 activities and 453 atomic actions in 1,752 synchronized sequences and 5,700 videos in total. We split the dataset into 1,388 train sequences and two test splits containing 198 and 166 sequences each. Each sequence has a high-level activity category. We annotated atomic actions in each of these videos by

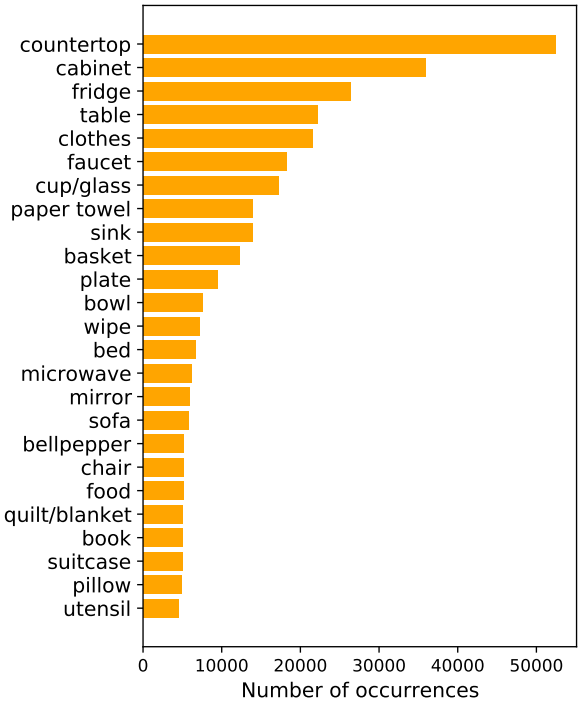


Figure 4.3: Distribution of object classes (top 25).

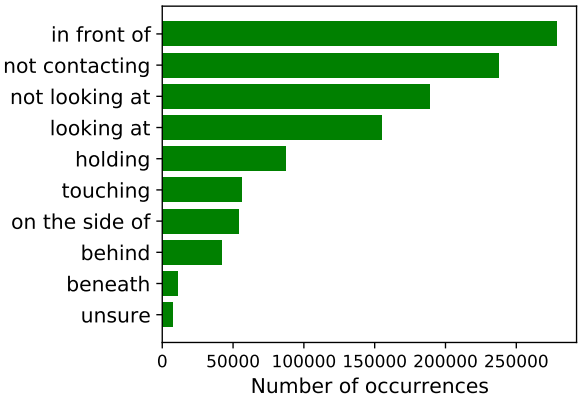


Figure 4.4: Distribution of relationship classes (top 10).

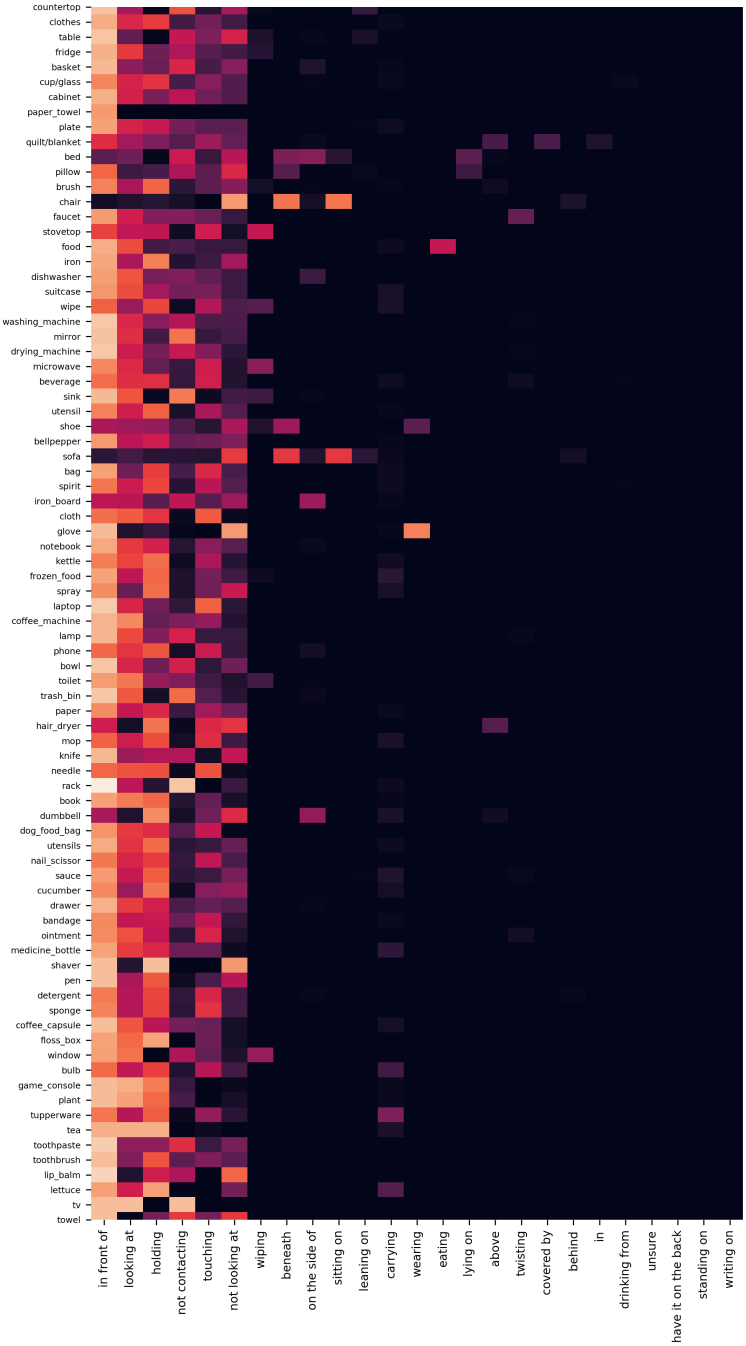


Figure 4.5: The co-occurrence statistics for objects and relationships in Home Action Genome.

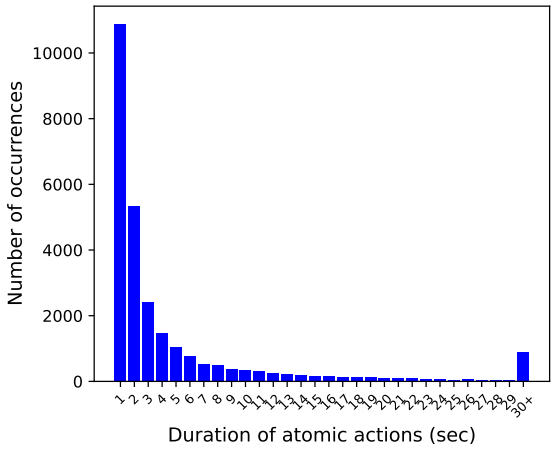


Figure 4.6: Distribution of duration of atomic actions.

providing the start and end frames and the category of the atomic action. There are 20,039 training, 2,062, and 2,468 atomic action sequences in the three splits mentioned above respectively. For scene graphs, we annotate one third-person view video in each synchronized sequence by providing bounding boxes of the subject and the object along with the relationship between them. There are 86 object classes (excluding "person"), and 29 relationship classes in the dataset. Overall, there are annotations of 497,534 bounding boxes and 583,481 relationships. .

The duration of atomic actions in HOMAGE is often short in time: there are about 60% of the atomic actions under 2 seconds and 80% under 5 seconds. For scene graphs, some of the most common objects are "countertop," "clothes," and "table"; and the most common relationships include "in front of," "looking at," and "holding." More details on the statistics are available in the supplement.

For the spatio-temporal scene graph, Figure 4.3 shows the most frequent object classes and Figure 4.4 shows the most frequent object relationships. Figure 4.5 shows the joint distribution of object classes and relationships. Figure 4.6 shows the distribution of the durations of atomic actions. To encompass activities in the living space, the types of activities in this dataset were determined by referring to the American Time Use Survey (ATUS), which is a survey of time at home allowing researchers to look at how much time people spend doing different activities. As there are several existing references defining atomic action for daily activities, we borrow definitions from datasets such as Charades [72], EPIC-KITCHEN [10] and Action Genome [32].

[1] We here refer to the "task classes" in [34]

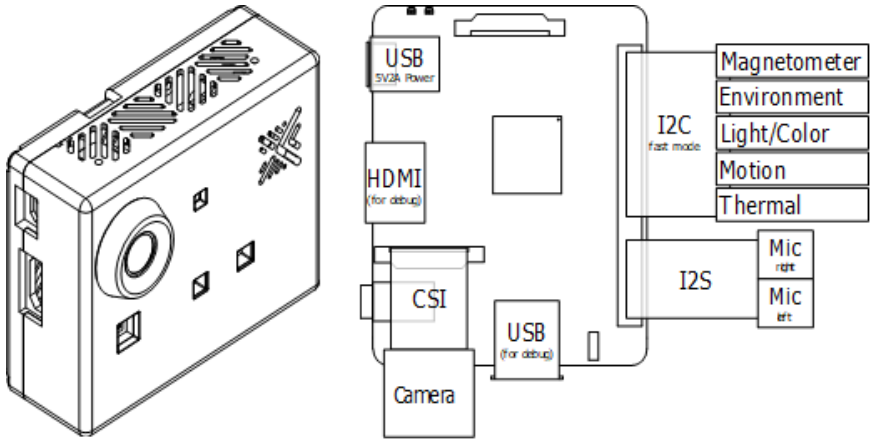


Figure 4.7: Multi-modal sensor kit used in data collection.

4.3.5 Relevance of Modalities.

In this paper, we only study the effect of modalities of interest to the vision community; however, HOMAGE provides rich sensor information which could be useful for privacy-aware recognition. Modalities such as angular velocity, acceleration, and geomagnetic sensors can be used to extract motion information in ego-view, and environmental sensors, e.g., temperature and humidity can capture changes in the scene before and after an activity. Thermal sensors can extract people or heat sources (e.g., extracting heat sources can be useful for recognition in places such as kitchens), and human presence and light sensors can determine the presence of people without using visual cues. Although not explored in detail in this paper, future sensor-fusion work can exploit these other modalities as well.

Sensors and Modalities

We build multi-modal sensor kits for data collection as shown in Figure 4.7. This kit assists the creation of the multi-modal dataset by dramatically simplifying the data collection process through simple recording and timing synchronization. The data from all viewpoints are collected by these sensor-kits. Figure 4.8 shows the photo of the multi-modal sensor mounted on the head of a subject participant.

The audio and video data from the sensor is saved to a video file, and the sensor data is saved in the same file as additional tracks. By using lossless codecs like the Free Lossless Audio Codec (FLAC) or WavPack, we can save the sensor data with high fidelity. Both codecs support multi-channel audio in 8-32 bit integer format at frequencies as low as 1Hz. Sensor data is acquired over I2C with constant timing adjustments to maintain synchronization with audio and video.

HOMAGE contains 12 modalities with multiple viewpoints. Specifically, the infrared data is obtained by the Grid-EYE 8x8 pixel infrared array sensor. The RGB light data is obtained by a photodiode array sensor that provides an RGB spectral response with IR blocking filter. The sensor kit also includes an ambient light

Figure 4.8: The sensor, mounted on the participant's head.

sensor that combines a broadband photodiode and an infrared-responding photodiode on a single CMOS-integrated circuit to provide ambient light data. The human presence sensor is a 4-channel nondispersive infrared (NDIR) sensor. The magnetic field data is acquired from a magnetometer in the sensor kit.

Data Synchronization

When storing video, audio, and sensor data together, each data stream is stored in a container by multiplexing the streams. We use H264 for the video stream, and FLAC (Free Lossless Audio Codec) for audio and sensor data.

To synchronize the sensor data, A 60Hz, fixed-length time-division multiplexing scheduler is used to query the sensors over the inter-integrated circuit (I2C) bus. The scheduler monitors the drift between expected and actual query times and adjusts its timing on the fly to achieve sub-millisecond accuracy on average. Sensor data are timestamped and passed to the main thread and encoded into its respective track immediately to guarantee synchronization.

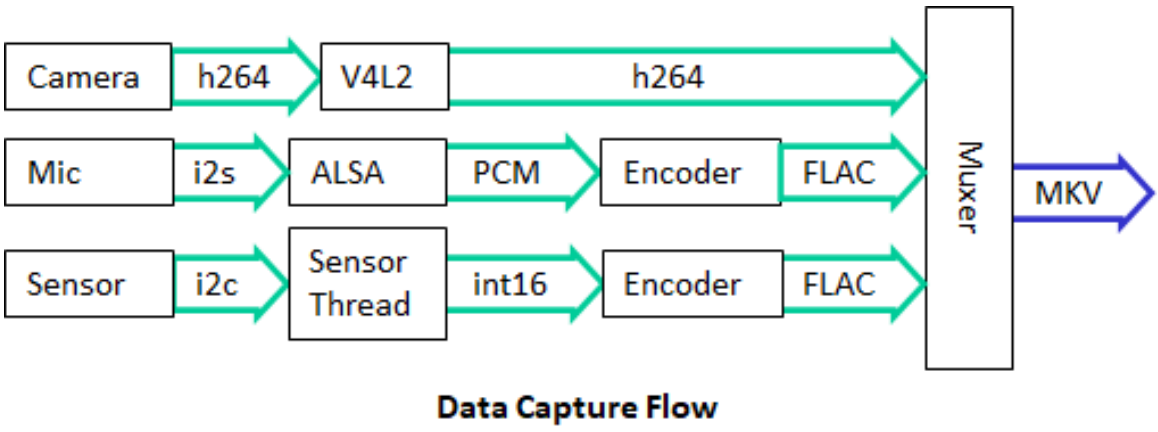


Figure 4.9: The flow chart of data collection.

Chapter 5

Cooperative Compositional Understanding

5.1 Introduction

In the previous section, we introduce a new benchmark for action recognition, Home Action Genome (HOMAGE), that includes multi-modal synchronized videos from multiple viewpoints along with hierarchical action and atomic-action labels. Actions in homes are challenging as we deal with long-term actions, interactions with objects, and frequent occlusions. Having multiple views and sensors to handle occlusions and scene graph information to capture object interaction allows us to tackle these complexities. In addition, synchronous videos provide implicit alignment that facilitates multi-modal training. Additionally, access to sensor information enables future research in privacy-aware recognition where we avoid audio-visual modalities. HOMAGE also provides temporal annotations of high-level activity and low-level atomic action supplemented with spatio-temporal scene-graphs. Annotations regarding interaction of objects within actions and atomic actions within high-level actions enable research in explainable video understanding, early action prediction, and long-range action recognition.

For this new benchmark, we introduce a novel method to perform simultaneous co-training with multiple modalities (RGB, audio, and annotations of scene composition) and viewpoints that enable the learning of rich video representations. Training involves a co-training strategy that leverages information from all views and modalities to build the representation space. During inference, we set up different experiments and observe improved action recognition performance even when only a single modality is used, which suggests training on HOMAGE improves performance with no need for other modalities during inference. In this paper, we explore audio-visual data (of interest to the vision community). Future sensor-fusion work can further exploit other modalities we release (e.g., for privacy-preserving studies).

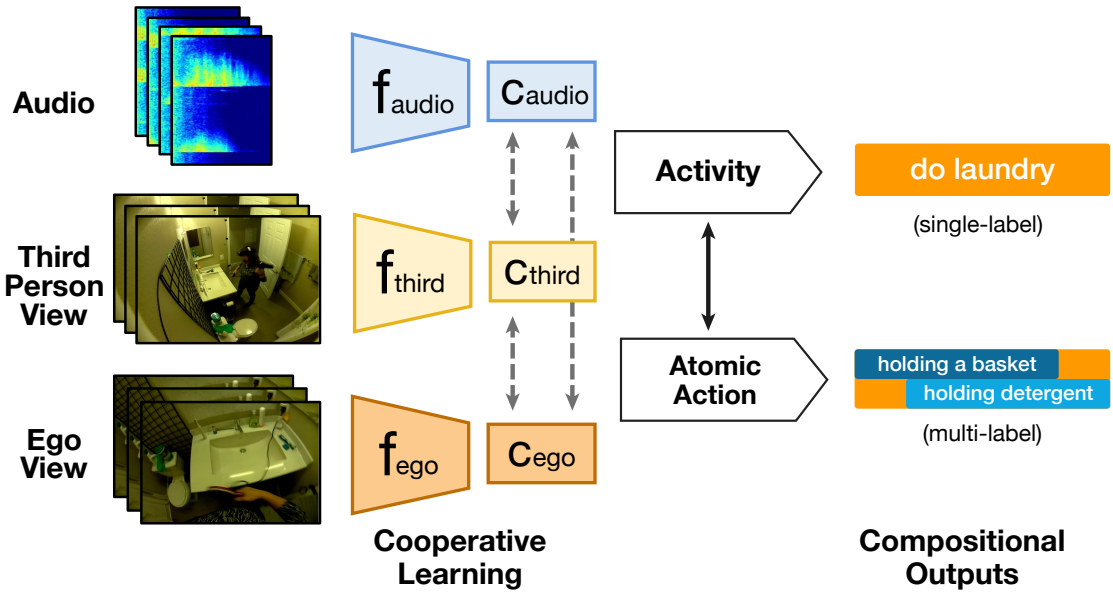


Figure 5.1: Given an activity instance (e.g., 'do laundry') and corresponding multiple views, we compute features using modality-specific deep encoders (f modules). Different modalities may capture different semantic information regarding the action. *Cooperatively* training all modalities together allows us to see improved performance. We utilize training using both video-level and atomic action labels to allow both the videos and atomic actions to benefit from the *compositional* interactions between the two. As discussed in the results, we see significantly improved performance when using the above components together.

HOMAGE aims to unify various aspects and challenges of action recognition, specifically targeting multi-modal and compositional perception for home actions. Moreover, the presence of a large number of modalities in our dataset encourages research in areas such as privacy-aware recognition and sensor-fusion. We discussed our contributions related to our dataset in the last section. In this section, we propose a novel learning framework (CCAU) that leverages multiple modalities and hierarchical action labels and improves the performance of the baselines trained on each individual modality. We demonstrate the benefits of our approach with an improvement of +6.4% using only ego-view during inference.

5.2 Cooperative Compositional Action Understanding

We discuss the benefits of HOMAGE and propose our approach Cooperative Compositional Action Understanding (CCAU) allowing us to exploit the rich annotations present in the dataset for improved action understanding. We discuss how CCAU employs simultaneous cooperative training with multiple modalities to improve the model's understanding of actions and the associated atomic-actions. We start by discussing a few preliminaries and proceed to discuss different components of our model. Note that "modalities" refer to both different camera views, as well as, modes such as images, audio, and scene graphs.

5.2.1 Preliminaries

A video V is a sequence of T frames with resolution $H \times W$ and C channels, $\{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_T\}$, where $\mathbf{i}_t \in \mathbb{R}^{H \times W \times C}$. Assume $T = N * K$, where N is the number of blocks and K denotes the number of frames per block. We partition a video clip V into N disjoint blocks $V = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_j \in \mathbb{R}^{K \times H \times W \times C}$ and a deep encoder $f(\cdot)$ transforms each input block x_j into its latent representation $z_j = f(x_j)$. An aggregation function, $g(\cdot)$ takes a sequence $\{z_1, z_2, \ldots, z_j\}$ as input and generates a context representation $c_j = g(z_1, z_2, \ldots, z_j)$. In our setup, $z_j \in \mathbb{R}^{H' \times W' \times D}$ and $c_j \in \mathbb{R}^D$. D represents the embedding size and H', W' represent down-sampled resolutions as different regions in z_j represent features for different spatial locations. We define $c = F(V)$, where $F(\cdot) = g(f(\cdot))$. In our experiments, $H' = 4, W' = 4, D = 256$. The computed representations are then utilized in order to perform per-block classification to generate the necessary predictions, e.g., activity label or atomic-action label. For multiple modalities, we define $c_m = F_m(V_m)$, where V_m, c_m and F_m represent the video input, context feature and composite encoder for modality m, respectively.

RGB Videos with Multiple Viewpoints.

An interesting aspect of HOMAGE is the presence of multiple viewpoints, specifically, a single ego-centric viewpoint and numerous third-person views. For simplicity, we treat these multiple viewpoints as two separate modalities, i.e., ego-view and third-person view. Each of these modalities has a dedicated encoder to generate clip-level features.

Audio.

Along with having multiple camera viewpoints, we also have associated audio clips for each viewpoint. For simplicity, we only use the audio associated with the ego-centric view. For each audio clip, we generate the associated log-mel spectrogram [1] and treat it as an image input. Following numerous other works [2, 41], we utilize a VGG19 backbone to generate a representation for the passed-in spectrogram.

Scene Graph.

A scene graph in a given frame G contains a set of objects $O = \{o_1, o_2, ...\}$ and a set of relationships $R = \{r_1, r_2, ...\}$. Each object o_j contains an object ID, bounding box coordinates of the object, and object category. Each relationship r_j contains the object IDs for both the subject and the object of the relationship, as well as the category of the relationship.

5.2.2 Multi-Modal Cooperative Learning

As discussed earlier, we define $c_m = F_m(V_m)$, where V_m, c_m and F_m represent the input, context feature, and composite encoder for modality m, respectively. We simultaneously train encoders for each modality while ensuring that the views improve with cooperation. Such a training regime allows us to observe improved performance during inference even when using a single modality.

Intuitively, we expect different modalities to impart complementary information to other modalities during training. This can be similar to existing approaches such as student-teacher frameworks or knowledge distillation [26, 40]. However, as we demonstrate in the experiments section, CCAU manages to learn better representations. We argue this is because the unidirectional formulation of student/teacher does not suit such setups as different modalities serve as a collective cohort of students as opposed to one of them being significantly dominant compared to others. CCAU utilizes contrastive multi-modal losses to promote cooperation between the learners.

Noise Contrastive Estimation (NCE) [18, 57, 60] constructs a binary classification task where a classifier is fed with real and noisy samples with the training objective to distinguish them. We utilize a simple task of performing alignment between different modes m, m'. The task becomes choosing the correct in-sync instance amongst multiple noisy instances. Similar to [80], we use an NCE loss over our feature embeddings c described in Eq. (5.1). c_i^m represents the feature embedding for the m^{th} modality's i^{th} temporal block. This effectively becomes a cross-entropy loss distinguishing one positive pair from all the negative pairs present in a video. In a batch setting with multiple video clips, it is possible to have more inter-clip negative pairs. The objective function for a single pair of modalities will hence be:

$$\mathcal{L}_{align}^{m,m'} = -\sum_i \left(\log \frac{\exp(c_i^m \cdot c_i^{m'})}{\sum_j \exp(c_i^m \cdot c_j^{m'})} \right). \tag{5.1}$$

To extend this to multiple views, we utilize the same objective for all pairs and simultaneously optimize:

$$\mathcal{L}_{align} = \sum_{m,m'} \mathcal{L}_{align}^{m,m'}.$$

Self-supervised attention [61] has been shown to be useful to auto-learn associations between different modalities. We model attention by predicting importance weights over the grid. We predict $H' \times W'$ values $\alpha_{i,j}$ representing weights of each feature corresponding to spatial location (i, j). Given feature c of shape $D \times H' \times W'$, we extract c_{agg} from it as given in Eq. (5.2). Where τ refers to the temperature. Further details are provided in the appendix.

$$c_{agg} = \sum_{i,j} p_{i,j} \cdot c_{i,j} \ \ , \ \ p_{i,j} = \frac{\exp(\alpha_{i,j} \ / \ \tau)}{\sum_{a,b} \exp(\alpha_{a,b} \ / \ \tau)} \tag{5.2}$$

5.2.3 Compositional Action Recognition

In addition to the multi-modal nature of HOMAGE, another one of its differentiating factors is having fine-grained atomic-action labels along with video-level action labels. The compositional nature of atomic-actions is useful in determining both the overall activity as well as learning relationships between atomic-actions and high-level actions.

We leverage the compositionality of atomic-actions and activities in CCAU by simultaneously utilizing both activity and atomic action level labels in our learning task. The intuition being our model will be able to learn the composition and relationships between atomic-actions and activities improving its understanding. We utilize the contextual features c in order to predict class labels for both video and atomic-action classes. The video action prediction task is a standard one-hot classification task, while we formulate the atomic-action prediction task as multi-target classification. We represent their corresponding losses as $\mathcal{L}_{video} = \mathcal{L}_v$ and $\mathcal{L}_{atomic} = \mathcal{L}_a$. The overall compositional loss is represented by $\mathcal{L}_{composition} = \mathcal{L}_c$.

We explore two variants to define $\mathcal{L}_c$. The first involves manually chosen hyper-parameters modulating each component, i.e., $\mathcal{L}_c = \mathcal{L}_v + \lambda \mathcal{L}_a$. The second automatically learns the appropriate multi-task weights [37]. The numbers reported in the paper represent use the first approach with $\lambda = 10$. For details refer to the appendix.

5.2.4 Self-Supervised Pre-Training

Our base backbone remains similar to the one we discuss in the main paper and the overall approach is inspired by [20]. To summarize, an aggregation function, $g(\cdot)$ takes a sequence $\{z_1, z_2, \ldots, z_j\}$ as input and generates a context representation $c_j = g(z_1, z_2, \ldots, z_j)$. In our setup, $z_j \in \mathbb{R}^{H' \times W' \times D}$ and $c_j \in \mathbb{R}^D$. D represents the embedding size and H', W' represent down-sampled resolutions as different regions in z_j represent features for different spatial locations. We define $z'_j = Pool(z_j)$ where $z'_j \in \mathbb{R}^D$ and $c = F(V)$ where $F(\cdot) = g(f(\cdot))$. In our experiments, $H' = 4, W' = 4, D = 256$.

To learn effective representations, we create a prediction task involving predicting z of future blocks similar to [20]. In the ideal scenario, the task should force our model to capture all the necessary contextual semantics in c_t and all frame-level semantics in z_t. We define $\phi(\cdot)$ which takes as input c_t and predicts the latent state of the future frames. The formulation is given in Eq. (5.3). Fig. 3.5 provides a compact visual representation of the learning framework.

$$\begin{aligned} \widetilde{z}_{t+1} &= \phi(c_t), \\ \widetilde{z}_{t+1} &= \phi(g(z_1, z_2, \ldots, z_t)), \\ \widetilde{z}_{t+2} &= \phi(g(z_1, z_2, \ldots, z_t, \widetilde{z}_{t+1})), \end{aligned} \tag{5.3}$$

where $\phi(\cdot)$ takes c_t as input and predicts the latent state of the future frames. We then utilize the predicted $\widetilde{z}_{t+1}$ to compute $\widetilde{c}_{t+1}$. We can repeat this for as many steps as we want, in our experiments we restrict ourselves to predict till 3 steps in to the future.

Note that we use the predicted $\widetilde{z}_{t+1}$ while predicting $\widetilde{z}_{t+2}$ to force the model to capture long-range semantics. We can repeat this for a varying number of steps, although the difficulty increases tremendously as the number of steps increases as seen in [20]. In our experiments, we predict the next three blocks using the first five blocks.

5.3 Experiments

We discussed the rich annotations in Home Action Genome (HOMAGE) that allows us to explore multiple aspects previously not possible due to the lack of such datasets. CCAU utilizes cooperative and compositional learning to learn improved representations for action understanding. Co-training with other modalities such as audio imparts additional structure and knowledge to individual modalities, also leading to improved single-view performance. We design and discuss multiple quantitative experiments to verify the validity of our claims. We also conduct qualitative experiments to gain deeper insights into our approach. In this section, we briefly go over our experiment framework. Additional details are provided in the appendix.

5.3.1 Dataset

HOMAGE provides a rich source of videos of human actions and multiple synchronized modalities representing complementary information about the action sequence. The hierarchical annotations in the dataset allow us to understand relationships between atomic actions and how they interact with each other to create higher-order actions.

5.3.2 Dataset Statistics

There are three splits in our dataset; 1,388 train segments and two test splits containing 198 and 166 segments each. Each segment includes synchronized videos from multiple views: one ego-view video and multiple third-person views, resulting in 1,752 segments and 5,700 videos in total. Each segment has a high-level activity category. Atomic actions are annotated in each of these videos. There are 20,039 training, 2,062 and 2,468 atomic action segments in the train split and the two test splits respectively, resulting in 24,569 atomic action segments and 78,261 videos in total. For each atomic action, we annotate the start and end frame, and the atomic action category. Note that, unlike activity labels, each atomic action segment can be assigned with multiple atomic action labels due to overlapping or the hierarchical nature of the ontology.

5.3.3 Modalities

The dataset contains synchronized multi-view videos grouped by action segments. Each segment contains an ego-view video and multiple third-person view videos. On average, there are more than 3 views per action segment. For both training and testing, we treat ego-view as one modality and all third-person view videos as another. Another rich source of information is the audio data and annotated scene graphs. The presence of complementary information, not found in RGB videos allows us to learn improved representations for our actions.

5.3.4 Implementation Details

Following our design discussed earlier to allow inference using individual modalities, we use separate encoders for each. We use different designs as mentioned in Section 5.2.1.

5.3.5 Images

In all of our experiments, we treat ego-view as one modality and all third-person view videos as another. We resize each input frame to the size of 128x128. We employ a 3D-ResNet similar to [24] as the encoder $f(\cdot)$. Following [20], we only expand the convolutional kernels present in the last two residual blocks to be 3D ones and use 3D-ResNet18 for our experiments, denoted as ResNet18. A weak aggregation function $g(\cdot)$ is used to learn a strong encoder $f(\cdot)$. Specifically, we use a one-layer Convolutional Gated Recurrent Unit (ConvGRU) with kernel size (1, 1) as $g(\cdot)$. The weights are shared amongst all spatial positions in the feature map. This design allows the aggregation function to propagate features in the temporal axis.

We use a dropout [27] with $p = 0.1$ to compute the hidden state at each time step. A shallow two-layer perceptron is used as the predictive function $\phi(\cdot)$. Recall $z'_j = Pool(z_j)$ where $z'_j \in R_D$. We utilize stacked max pool layers as $Pool(\cdot)$. To construct blocks to pass to the network, we uniformly choose one out of every 3 frames. Then, they are grouped into 8 blocks containing 5 frames each. Since the videos are usually 30fps, each block roughly covers 0.5 seconds and 8 blocks sums to about 4 seconds worth of action. Given the 256D

Method	Audio	Ego	$3rd$ Person
Single Modality	28.5	31.3	21.8
Cooperative Ours	**33.3**	**37.7**	**24.7**
Static KD	28.5	32.3	21.8
Cooperative KD	32.1	32.1	23.5

Table 5.1: Video classification accuracy. *Cooperative Ours* outperforms the baselines. *Cooperative KD* performs better than its counterparts, further validating benefits of cooperative learning.

final representations, we pass this through fully connected layers to compute the final classification where we use a dropout of $p = 0.5$.

5.3.6 Audio

To process audio clips, we convert audio to MP3 format, compute log-mel spectrograms [1], and pass it through a VGG19-like convolutional architecture. We sample fixed intervals of the spectrogram image to represent the action clip. Similar to the image encoder, we have fully connected layers to perform classification.

5.3.7 Scene Graphs

We use ground-truth scene graphs to predict action labels. For each given frame, we fetch the scene graph that is the closest to this frame. We encode the scene graph as a matrix with dimensions representing the categories of objects and that of relationships. We flatten the matrix to get the scene graph input features. For the feature encoder, we use 3 layers of linear layers with ReLU and hidden dimensions of 256. We employ a dropout rate of 0.5 before the final fc layer.

5.4 Quantitative Results

In this section, we analyze various aspects of our proposed model. To objectively evaluate model performance, classification accuracy is utilized as a proxy for learned representation quality. Evaluation is performed on two different splits of HOMAGE. Although models have access to other modalities during training, this is not the case during inference. Therefore, evaluation only involves inference using individual modalities. However, we see an improvement despite this constraint due to co-training. We also study the improvement imparted through compositional learning with both high-level action and atomic-action labels.

5.4.1 Comparisons with Baselines

In this section, we investigate the effectiveness of cooperative multi-modal learning for action understanding. We study the impact of cooperative learning and compare the performance to knowledge distillation approaches.

Impact of Cooperation.

Our co-operative training approach hinges on the assumption that multi-modal information helps in improving overall representation quality. To verify our hypothesis, we study the performance of CCAU compared along with a few other comparable approaches. (1) *Single Modality Training (SM)* - Training of modalities independently (2) *Cooperative Ours Training (CT)* - Co-Training of all modalities and individual inference. Table 5.1 summarizes our results demonstrating a consistent improvement in performance across modalities.

Comparison with Knowledge Distillation.

Given the potential applicability of student-teacher approaches in this setting, we also study their performance compared to our approach. We study two variants. (1) *Static Knowledge Distillation (SKD)* - We transfer knowledge from other trained modalities into the ego-view encoder. (2) *Cooperative Knowledge Distillation (CKD)* - To isolate the effect of cooperation leading to improved performance, we also propose a cooperative version of knowledge distillation that allows all modalities to simultaneously improve (details in the appendix). Table 5.1 summarizes our results demonstrating the performance difference between these approaches. We notice a performance improvement when utilizing cooperative KD compared to the static variant. CT outperforms CKD even though both allow cooperation, due to the incorrect student-teacher hierarchy even with a symmetric knowledge distillation setup. CT allows cooperation in a softer manner without an implicit assumption of hierarchy.

Impact of Additional Modalities

We saw the benefits of Cooperative Training in the previous section and established the performance improvements accompanying training with multiple modalities. In this section, we look at the implications modalities have on performance by studying the impact of training with multiple modalities. We consider 1) *Training each modality separately*; 2) *Joint training of multi-camera views*, i.e., Ego and $3rd$ Person RGB video clips, and 3) *Joint training of multi-camera views with ego-centric audio clips*.

Activity Classification. Table 5.2 summarizes the results of our approach trained with different modalities. Compared with training with single views individually, co-training with the two video views and video + audio consistently improves the performance together with more modalities.

Atomic Action Classification. We also investigate the impact of cooperative training on multi-target classification for atomic actions. Table 5.3 summarizes our results. The Mean Average Precision scores for each modality are reported.

Method	Audio	Ego	$3rd$
Single Modality	28.5	31.3	21.8
Coop - Ego + $3rd$	-	35.1	23.5
Coop - Ego + $3rd$ + Aud	**33.3**	**37.7**	**24.7**

Table 5.2: Co-training encoders with different modalities on activity classification. We see a distinct performance improvement across modalities as we co-train with increasing number of modes, possibly due to the presence of rich complementary information.

Method	Audio	Ego	$3rd$ Person
Single Modality	7.0	20.5	11.7
Cooperative	**13.2**	**28.5**	**15.3**

Table 5.3: Effect of co-training encoders with different modalities on atomic action classification. The numbers reported are support weighted mAP scores.

5.4.2 Cooperative Compositional Learning

We analyze the role of both our proposed soft attention module and CCAU's compositional learning framework.

Impact of Co-training with Attention.

Table 5.4 summarizes the results of the cross-modality co-training experiment with and without attention module. With attention, the model yields better accuracy on the video modalities compared with its counterpart. The model can implicitly learn localization and correspondence between views to form representations with view-invariant information.

Impact of Compositional Learning.

Our compositional learning framework hinges on the assumption that simultaneously learning both activity labels and atomic action labels leads to improved performance. To verify this hypothesis, we compare different variants such as (1) train with activity labels, (2) train with atomic-action labels, (3) train with

Method	Ego	$3rd$ Person
Cooperative	32.5	19.1
Cooperative with Attention	**34.8**	**20.8**

Table 5.4: Effect of co-training encoders using the proposed attention module. We see a consistent performance improvement across both modalities. The $3rd$ person mode benefits as attention allows potential localization of the region of interest - despite the lack of dense associations between the ego and $3rd$ person view.

Method	Acc			mAP		
	Audio	Ego	$3rd$ Person	Audio	Ego	$3rd$ Person
Cooperative - Activity	28.3	31.1	17.0	-	-	-
Cooperative - Atomic Actions	-	-	-	5.9	18.5	9.5
Compositional	23.5	32.1	16.2	16.4	26.3	12.2
Cooperative Compositional	**29.3**	**34.9**	**19.2**	**21.7**	**29.3**	**13.8**

Table 5.5: Effect of co-training encoders with images and audio on activity classification. We see a distinct performance improvement compared to the Ego, $3rd$ Person Co-Training case; due to the rich complementary information present in audio encoders. Missing numbers denote the model was not trained for the associated subtask. Results are averaged over the two test splits.

Method - Ego	Atomic Action - mAP			
	1 shot	5 shot	10 shot	20 shot
Single Modality	22.4	35.3	38.6	40.6
CCAU	**28.6**	**36.9**	**39.4**	**49.4**

Table 5.6: Compositional learning with few shot learning. With compositional action understanding, CCAU demonstrates much better generalizability than other baseline, showing the potential of co-learning with compositional labels in improving action understanding. Results are averaged over the two testing splits.

both activity and atomic actions without cooperation and (4) CCAU - cooperatively train with both video and atomic actions. In Table 5.5, we see a consistent improvement across both activity and atomic-action performance.

5.4.3 Few-Shot Compositional Action Learning

We have discussed the benefits of our cooperative and compositional approach. Intuitively, predicting activities should be easier if we have an idea of the atomic-actions composing the higher-order action. We now showcase the ability and potential of CCAU to generalize to rare actions.

Setup.

In our few-shot action recognition experiments, we split the 75 action classes into a base set of 60 classes and a novel set of 15 classes. We use CCAU as our feature extractor. Note that we do not finetune the backbone. Next, we train each model with only k examples from each novel class, where k = 1, 5, 10, 20. Finally, we evaluate the trained models on all examples of novel classes in the validation set.

Results.

We report few-shot experiment performance in Table 5.6. CCAU improves the single modality baseline on all 1, 5, 10, 20-shot experiments. Furthermore, CCAU shows a +6.2% 1-shot and +8.8% 20-shot mAP

Method	Ego-View	3^{rd} Person
SV	31.8	21.8
SS + SV	**33.1**	**24.8**

Table 5.7: Effect of self-supervised pre-training on atomic action classification. We see considerable performance improvements when initializing our model with pre-training using multi-modal self supervision. This results in distinctively improved performance compared to random initialization as we're able to utilize structural information naturally present in the examples. This demonstrates the additional possibility of utilizing Home Action Genome in order to evaluate multi modal self-supervision approaches.

improvement.

5.4.4 Additional Results

In order to study the effectiveness of our approach, we perform additional experiments targeting different features of our proposed model.

Self-Supervised Pre-training

To study the value of multiple viewpoints of the video data, we perform pre-training with the above learning framework weights to get a self-supervised initialization for our experiment. We first train our model in the self-supervised setting for 500 epochs. We use the pre-trained weights to initialize the ego-view and third-person view encoders and train with supervision loss to the same number of epochs as the randomly initialized baseline. Note that in the supervision phase, each modality is trained separately and no cross-modality loss is used. Table 5.7 shows that cooperative learning with different modalities results in distinctively improved performance compared to random initialization as we are able to utilize structural information naturally present in the examples. We also observe that the model with self-supervised pre-training converges faster than the baseline. This demonstrates the additional possibility of utilizing Home Action Genome to evaluate multi-modal self-supervision approaches.

Baseline with Oracle Scene Graphs

We provide a baseline for human action classification using oracle scene graphs. This experiment gives a rough reference of the upper bound of action inference using spatio-temporal information.

We represent the ground-truth scene graph input as a matrix M of size $n_{obj} \times n_{rel}$, with n_{obj} and n_{rel} be the number of object and relationship categories, respectively, initialized to be filled with 0. We encode a relationship with object category s, and relationship category r by setting $M[s, r]$ to be 1. The input representation is then flattened and fed into an MLP-based encoder.

Table 5.8 shows the performance of activity classification using ground-truth scene graphs, with the encoding scheme described above. We observe that the modality of the ground-truth scene graph is very informative compared with the other modalities, highlighting the potential for scene graph prediction on human action understanding.

Acc1	Acc3
76.0	91.7

Table 5.8: Classification of activities using ground-truth scene graphs. Results are averaged over the two test splits.

Multi-Task Loss

As discussed in Section 5.2.3, we utilize two variants for our multi-task losses. The first is an equally weighed variant where both $\mathcal{L}_a$ and $\mathcal{L}_v$ have the same weights, while the other is similar to the one proposed in [37] utilizing task-dependent uncertainty to automatically weigh losses. The loss is defined as:

$$\mathcal{L}_c = \mathcal{L}_v \,/\, \sigma_v^2 + \mathcal{L}_a \,/\, \sigma_a^2 + \log(\sigma_v.\sigma_a) \tag{5.4}$$

Where σ_i refers to the task dependent uncertainty (aleatoric homoscedastic). Although the latter has shown improved results in numerous settings, we noticed that it led to slower convergence and the performance improvements were not consistent across modalities. For this reason, all results reported utilize the simple equally weighted multi-task loss.

Learning Attention

As mentioned in the main text, we also explore the usage of an attention module that allows auto-learning of associations between different modalities similar to [61] which do it for audio and visual modalities. We setup attention in a slightly different manner by predicting weights over the grid. Recall that our features are arranged in a grid of shape $H' \times W'$. We predict $H' \times W'$ values $\alpha_{i,j}$ representing the weight of each feature corresponding to spatial location (i, j). Given an original context c of shape $D \times H' \times W'$, we extract c_{agg} from it as given in Eq. (5.2). Note that we generate attention weights for each pair of modalities to capture the associations between them.

In our experiments, we did not notice any differences between choosing various values of temperature as it seems the network modulated the learned α's accordingly. p's are utilized to infer regions of interest, as cells with higher p correspond to relevant portions of the modalities. Another thing worth noting is that this attention module is only used in conjunction with image modalities, as we found attention over an audio spectrogram was not directly interpretable in the traditional sense.

Figure 5.2: Visual results for multi-modal attention between ego-centric and third person view. We show four instances where the left image refers to the third person view, while the right shows the predicted attention weights (White represents higher importance for attention). As we can see, CCAU is loosely able to predict areas of interest using our proposed self-supervised losses.

Knowledge Distillation

We discuss Knowledge Distillation briefly in the main text as one of the important baselines in Section 5.3.1. The framework we used is similar to the famously used one proposed in [26]. Without going into details, the overall loss is given in Eq. (5.5).

$$\mathcal{L}_{kd} = \alpha \cdot \mathcal{H}(y, \sigma(zs)) + \beta \cdot \mathcal{H}(\sigma(zt, \tau), \sigma(zs, \tau)) \tag{5.5}$$

Eq. (5.5) is an instance of matching logit distributions leading to the distillation of knowledge from the teacher to the student. Where H represents the cross-entropy loss, τ represents the temperature. zs and zt are outputs for the student and teacher, respectively.

For multiple modalities, the loss is just repeated multiple times for each modality. For our experiments we use $\alpha = 1$ and $\beta = 0.1$. We choose $\tau = 2.5$ as the models are similar in capacity. We also experiment with two variants i.e. Static and Cooperative Knowledge Distillation. The difference being Static KD involves static teachers while the cooperative variants allow all modalities to serve as both students and teachers.

5.5 Qualitative Results

One of the motivating factors behind CCAU was the benefits of co-training different encoders together to gain higher-order perspectives provided through different modalities. We observe the learned structure across

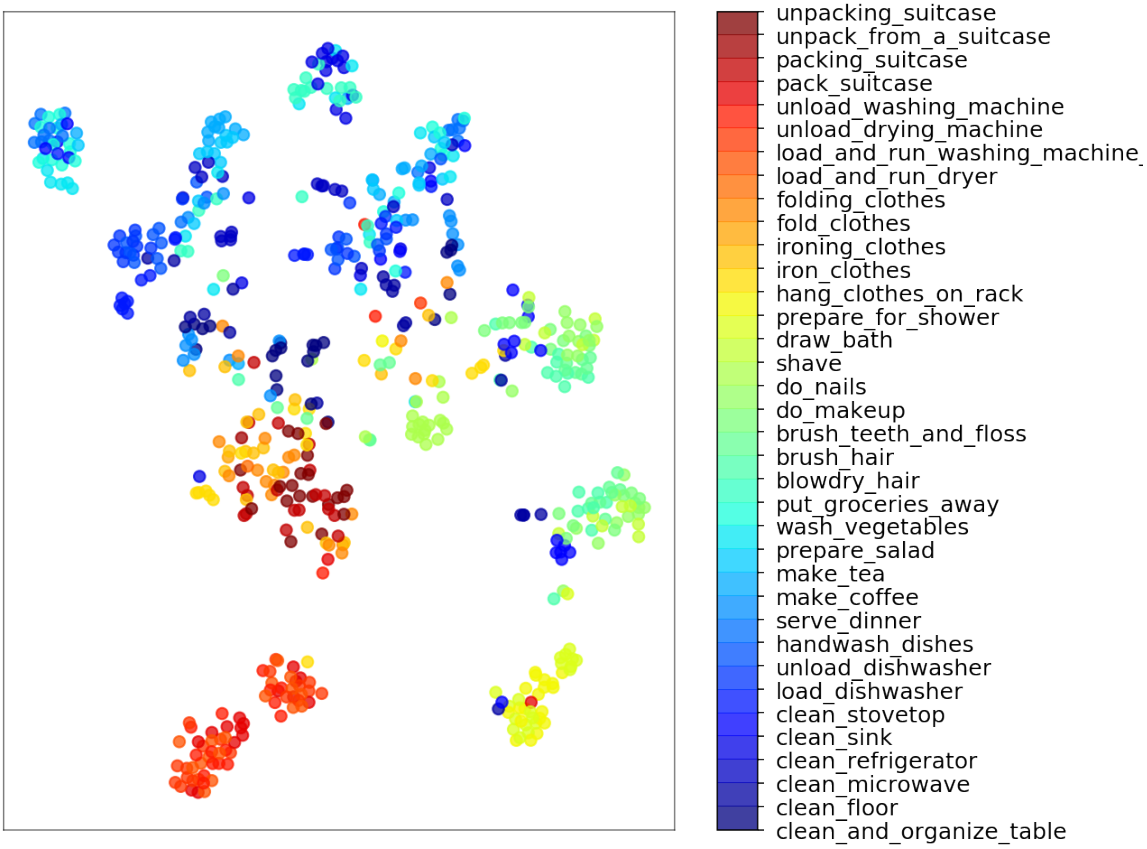


Figure 5.3: t-SNE visualization of Ego-View features from *CCAU* trained with ego, $3rd$ and audio modalities. The color mapping represents the relationships between the action classes, e.g., Red: Clothes; Green: Grooming; Blue: Kitchen. CCAU is able to learn meaningful clusters by utilizing compositional information.

modalities results in the emergence of higher-order semantics without additional supervision, e.g., sensible class relationships and good feature representations. Jointly training with modalities gives rise to better representations and byproducts such as localization of visual regions of interest.

5.5.1 t-SNE Visualization.

We explore t-SNE visualizations of our learned representations. For clarity, only a few action classes are displayed. We loosely order the action classes according to their relationships; classes having similar colors are semantically similar. Fig. 5.3 summarizes our results.

5.5.2 Multi-Modal Localization.

A by-product of learning attention using contrastive losses is the ability to localize potential points of interest in images (details in the appendix).

5.6 Conclusion

We introduced Home Action Genome (HOMAGE), a human action recognition benchmark with multiple modalities and viewpoints with hierarchical activity and atomic action labels. We also proposed CCAU, a cooperative and compositional learning method to leverage information across multiple modalities along with action compositions in HOMAGE for better representation learning. Due to the nature of cooperative learning, CCAU allows inference on individual modalities where no privileged information and other modalities are available. We demonstrated the benefits of learning atomic-actions compositions leading to significantly improved results in a few-shot learning setting.

With rich multi-modal data and compositional annotations, HOMAGE facilitates research in subfields such as multi–modal action recognition and localization, explainable action understanding, and reasoning with spatio-temporal scene graphs. We hope HOMAGE promotes research in multi-modal cooperative learning and action understanding using compositions for richer feature representations in human action recognition as well as raises interest in generalizable video understanding.

(1) Action Localization - Using multiple modalities to perform localization is gaining traction recently. We hope the presence of rich annotations in HOMAGE allows advancement in multi-modal localization-related research. (2) Explainable Action Understanding - Explainable visual models are receiving increased attention recently. Saliency prediction is a popular relevant approach. The presence of actor-object interest along with multiple modalities leads to robustness (e.g. to occlusions) and can be used to further improve existing explainable models. (3) Multi-Modal Understanding - There has been a surge in approaches exploiting multiple modalities of data while reducing the amount of supervision required.

Chapter 6

Conclusions and Future Directions

6.1 Conclusion

The underlying motivation in the ideas presented in this thesis has been reducing the need for supervision to learn and understand videos. Specifically, we target this task by exploring the combination of self-supervised learning and multi-modal learning and their combinations. We started off by introducing the benefits of utilizing multi-modal cues in a self supervised setting and used the cues to propose a cooperative contrastive learning approach resulting in better learnt representations. Our approach tackles a current drawback of self supervised methods where they incorrectly encourage semantically similar instance representations to be far apart simply because they are from different instances. We exploit multi modal cues to alleviate this issue in our proposed approach, leading to significant gains.

We then shift our focus to utilizing multiple views of our data to improve action understanding. We discuss our proposed dataset, Home Action Genome, which provides multiple modalities, camera views and other rich annotations such as hierarchical atomic action labels. Having such rich annotations during training allows us to impart additional knowledge into our models improving semantic understanding in videos. We discuss the features of our dataset and touch upon the benefits of having such a richly annotated dataset.

Once we introduce the specifics of Home Action Genome, we proceed to discuss our proposed approach, Cooperative Compositional Action Understanding, which allows us to use these modalities, camera views and annotations during training to improve the learnt video representations. We discuss numerous experiments highlighting different aspects of our approach and how it leads to improved performance even while performing inference without having access to other modalities.

Through these discussions, we demonstrate how utilizing readily available information in the form of multiple modalities and injecting intuition and information from real world through atomic action compositions, allows us to reduce the need for supervision and still obtain competitive performance. We also demonstrate the potential of self-supervised learning when used in conjunction with such approaches. We hope this encourages research in this direction allowing future machine learning algorithms to learn with less data and

become robust using the implicit information present in existing datasets.

6.2 Future Directions and Applications

We have explored self-supervised learning and multi-modal learning along with their intersection to aid action understanding in videos. We took a look at its implications on reducing the amount of supervision through numerous quantitative and qualitative results. We also explored using multiple views of our data during training to improve performance at inference time without having access to other modalities. Finally, we investigated how adding real world intuition and information through injecting atomic action compositions helps robustness and performance. There are, however, many more applications and experiments for which our approach and dataset can be used. We briefly discuss a few potential applications in this section.

6.2.1 Action Localization

Action localization has been a popular task for a while now and using multiple modalities to achieve it is gaining traction recently. As discussed earlier, different modalities provide complementary information which can be used to improve model performance consequently helping in action localization. We hope the presence of rich annotations in HOMAGE allows advancement in multi-modal localization-related research.

6.2.2 Action Alignment in Videos

Action alignment across videos is a useful task to infer associations and similarities between videos. We have demonstrated the potential of performing action alignment even in the absence of labelled data earlier. We hope our proposed approach encourages work in this direction.

6.2.3 Explainable Action Understanding

Explainable visual models are receiving increased attention recently. Saliency prediction is a popular relevant approach. The presence of actor-object interest along with multiple modalities leads to robustness (e.g. to occlusion) and can be used to further improve existing explainable models. The presence of intuitive annotations in HOMAGE can allow us to explore this direction.

6.2.4 Multi-modal Action Understanding

There has been a surge in approaches exploiting multiple modalities of data while reducing the amount of supervision required. The approaches discussed in this thesis i.e. our cooperative self supervised approach and our cooperative compositional approach designed for HOMAGE demonstrate this through numerous experiments and results. Having access to a rich dataset such as HOMAGE with numerous modalities can allow research in this direction.

6.2.5 Privacy Aware Action Understanding

We have seen that our cooperative training strategy allows us to observe improved action recognition performance even when only a single modality is used during inference, suggesting training on HOMAGE improves performance with no need for other modalities during inference. Although we have only explored audio-visual data in our experiments here, future sensor-fusion work can further exploit the other modalities we have released. Privacy-aware recognition has started gaining popularity where audio-visual modalities may be avoided. Access to sensor info allows future research to tackle this challenge effectively by improving performance of such modalities during inference by cooperatively training them together initially. Such a combination is not seen in existing datasets and we hope HOMAGE aids future researchers.

Bibliography

[1] Librosa mel spectrogram. `https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html`.

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. *arXiv preprint arXiv:2008.04237*, 2020.

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *arXiv preprint arXiv:1911.12667*, 2019.

[4] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[5] Ulf Blanke and Bernt Schiele. Remember and transfer what you have learned-recognizing composite activities based on activity spotting. In *International Symposium on Wearable Computers (ISWC) 2010*, pages 1–8. IEEE, 2010.

[6] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[7] João Carreira, Andrew Zisserman, and Quo Vadis. Action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733, 2018.

[8] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[11] Hanneke EM Den Ouden, Peter Kok, and Floris P De Lange. How prediction errors shape perception, attention, and motivation. *Frontiers in psychology*, 3:548, 2012.

[12] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017.

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

[14] 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.

[15] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

[16] Chuang Gan, Boqing Gong, Kun Liu, Hao Su, and Leonidas J Guibas. Geometry guided convolutional neural networks for self-supervised video representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5589–5597, 2018.

[17] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[18] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[19] Daniel S Hamermesh, Harley Frazis, and Jay Stewart. Data watch: The american time use survey. *Journal of Economic Perspectives*, 19(1):221–232, 2005.

[20] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Workshop on Large Scale Holistic Video Understanding, ICCV*, 2019.

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European Conference on Computer Vision*, 2020.

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

[23] Albert Haque, Arnold Milstein, and Li Fei-Fei. Illuminating the dark spaces of healthcare with ambient intelligence. *Nature*, 585(7824):193–202, 2020.

[24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[27] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors (2012). *arXiv preprint*

arXiv:1207.0580, 2012.

[28] Jakob Hohwy. *The predictive mind*. Oxford University Press, 2013.

[29] Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. *arXiv preprint arXiv:2003.06409*, 2020.

[30] Ahmad Jalal, Shaharyar Kamal, and Daijin Kim. A depth video-based human detection and activity recognition using multi-features and embedded hidden markov models for health care monitoring systems. *International Journal of Interactive Multimedia & Artificial Intelligence*, 4(4), 2017.

[31] Dinesh Jayaraman and Kristen Grauman. Slow and steady feature analysis: higher order temporal coherence in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3852–3861, 2016.

[32] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[33] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020.

[34] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–786, 2020.

[35] Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *ArXiv*, abs/1811.11387, 2018.

[36] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[37] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[38] 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.

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8658–8667, 2019.

[41] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018.

[42] 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, pages 780–787, 2014.

[43] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017.

[44] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017.

[45] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–14. IEEE, 2010.

[46] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

[47] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[48] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017.

[49] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[50] Mandy Lu, Kathleen Poston, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Kilian M Pohl, Juan Carlos Niebles, and Ehsan Adeli. Vision-based estimation of mds-updrs gait scores for assessing parkinson's disease motor severity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–647. Springer, 2020.

[51] Zelun Luo, Jun-Ting Hsieh, Niranjan Balachandar, Serena Yeung, Guido Pusiol, Jay Luxenberg, Grace Li, Li-Jia Li, N Lance Downing, Arnold Milstein, et al. Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring. *Machine Learning for Healthcare (MLHC)*, 2, 2018.

[52] Zelun Luo, Boya Peng, De-An Huang, Alexandre Alahi, and Li Fei-Fei. Unsupervised learning of long-term motion dynamics for videos. *CoRR*, abs/1701.01821, 2017.

[53] Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2020.

[54] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[55] 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 international conference on computer vision*, pages 2630–2640, 2019.

[56] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using

temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[57] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, pages 2265–2273, 2013.

[58] Bingbing Ni, Gang Wang, and Pierre Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1147–1153. IEEE, 2011.

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

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

[61] 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.

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

[63] Nishant Rai, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, and Juan Carlos Niebles. Cocon: Cooperative-contrastive learning, 2021.

[64] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding, 2021.

[65] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–771, 2018.

[66] Fahimeh Rezazadegan, Sareh Shirazi, Ben Upcrofit, and Michael Milford. Action recognition: From static datasets to moving robots. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3185–3191. IEEE, 2017.

[67] Mohammad Sabokrou, Mohammad Khalooei, and Ehsan Adeli. Self-supervised representation learning via neighborhood-relational encoding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8010–8019, 2019.

[68] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. Calibme: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 2594–2605, 2017.

[69] Nawid Sayed, Biagio Brattoli, and Björn Ommer. Cross and learn: Cross-modal self-supervision. *CoRR*, abs/1811.03879, 2018.

[70] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[71] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1-2):4–27, 2010.

[72] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 510–526. Springer, 2016.

[73] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016.

[74] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press.

[75] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[76] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[77] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.

[78] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011.

[79] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019.

[80] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019.

[81] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[82] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017.

[83] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating the future by watching unlabeled video. *CoRR*, abs/1504.08023, 2015.

[84] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *CoRR*, abs/1609.02612, 2016.

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

[86] 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*, pages

8052–8060, 2018.

[87] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[88] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[89] Jeffrey M Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological bulletin*, 127(1):3, 2001.

[90] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[91] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018.