Home Action Genome: Cooperative Compositional Action Understanding

Nishant Rai[1], Haofeng Chen[1], Jingwei Ji[1], Rishi Desai[1], Kazuki Kozuka[2], Shun Ishizaka[2],
Ehsan Adeli[1], Juan Carlos Niebles[1]
[1]Stanford University   [2]Panasonic Corporation
http://www.homeactiongenome.org

Abstract

Existing research on action recognition treats activities as monolithic events occurring in videos. Recently, the benefits of formulating actions as a combination of atomic-actions have shown promise in improving action understanding with the emergence of datasets containing such annotations, allowing us to learn representations capturing this information. However, there remains a lack of studies that extend action composition and leverage multiple view-points and multiple modalities of data for representation learning. To promote research in this direction, we introduce Home Action Genome (HOMAGE): a multi-view action dataset with multiple modalities and view-points supplemented with hierarchical activity and atomic action labels together with dense scene composition labels. Leveraging rich multi-modal and multi-view settings, we propose Cooperative Compositional Action Understanding (CCAU), a cooperative learning framework for hierarchical action recognition that is aware of compositional action elements. CCAU shows consistent performance improvements across all modalities. Furthermore, we demonstrate the utility of co-learning compositions in few-shot action recognition by achieving 28.6% mAP with just a single sample.

1. Introduction

Action understanding in videos is a critical task with various use-cases and real-world applications, from robotics [1, 2] and human-computer interaction [3] to healthcare [4, 5] and elderly behavior monitoring [6, 7]. 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
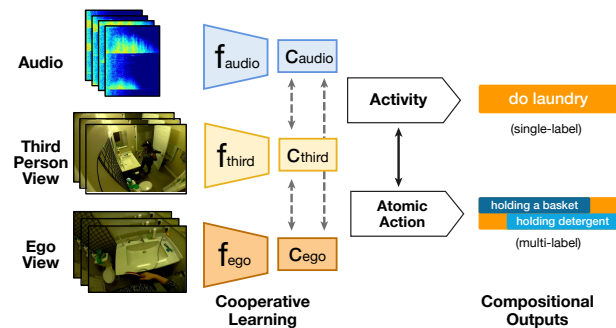

Figure 1: Given an activity instance (e.g., 'do laundry') and corresponding multiple views, we compute features using modality-specific deep encoders (f modules). Different modalities may capture different semantic information regarding the action. *Cooperatively* training all modalities together allows 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.

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 [8, 9, 10] 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 [11], Charades [12], ActivityNet [13], UCF101 [14]. Other datasets such as EPIC Kitchens [15] were used for ego-centric action recognition. Action Genome [10] focused on using scene information in action recognition,

while others [16] focused on hierarchical action modeling from events to low-level atomic actions. Several studies target learning from long instructional videos and release datasets [17, 18, 19, 20] 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 [21] and HumanEva [22]. In parallel, there have been numerous recent advances in contrastive and cooperative learning [23, 24] 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 [25, 26, 27] 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.

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.

2. Related Work

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 [25] 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. [28, 29] demonstrated that replacing 2D convolutions with 3D convolutions leads to further performance improvements. Recent approaches such as I3D [30] inflate a 2D convolutional network into 3D to benefit from the use of pre-trained models. 3D-ResNet [31] adds residual connections building a very deep 3D network leading to improved performance.

Related Datasets. MSR-Action3D [32] provides depth map sequences containing 20 actions of interactions with game consoles. [33, 34, 35, 36] use the Microsoft Kinect sensor to collect multi-modal action data with RGB and depth map sequences. NTU RGB+D [35] 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 [13, 33, 37, 21, 10] provide annotations for temporally localized actions. MMAct [37] 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 local-

ization and relationships between objects. LEMMA [21] 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 [10] is built upon the videos from Charades [38], 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.

Multi-Modal Learning. Multiple modalities of videos are rich sources of information for both supervised [25] and self-supervised learning [26, 27, 39]. [40, 27] 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 [25] have led to many competitive approaches, which demonstrate using even derivable modalities such as optical flow helps improve performance considerably. There have been approaches [41, 39, 27, 26] utilizing diverse modalities, sometimes derivable from one other, to learn better representations.

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.

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) [44]. 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

Figure 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.

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.

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.

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 3 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. [45] also shows this sampling scheme performs remarkably well.

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 se-

Dataset	Seq	hrs	Modalities	Views	HL	HL Classes	TL	TL Classes	TL Ins	SG
RGBD-HuDaAct [34]	1.19K	46	2	1	✓	12	-	-	-	-
UCF101 [14]	13K	27	1	1	✓	101	-	-	-	-
ActivityNet [13]	28K	648	1	1	✓	200	-	-	-	-
Kinetics-700 [11]	650K	1.79K	1	1	✓	700	-	-	-	-
AVA [42]	430	108	1	1	-	-	✓	80	1.58M	-
PKU-MMD [33]	1.08K	50	3	3	-	-	✓	51	20K	-
EPIC-Kitchens [15]	-	55	1	1	-	-	✓	125	39.6K	-
MMAct [37]	36K	-	6	5	-	-	✓	37	36.8K	-
Action Genome [10]	10K	82	1	1	-	-	✓	157	66.5K	✓
Breakfast [43]	-	77	1	1	✓	10	✓	48	-	-
LEMMA [21]	324	10.1	2	4	✓	15[1]	✓	863	11.8K	-
Ours	1.75K	25.4	12	2∼5	✓	75	✓	453	24.6K	✓

Table 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.
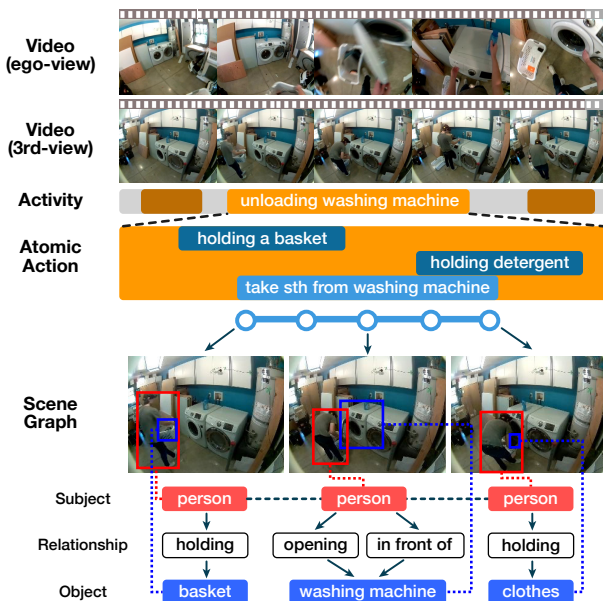

Figure 3: HOMAGE annotation pipeline: For every action, we uniformly sample 3 or 5 frames across the action and annotate the bounding boxes of the person performing the action along with the objects they interact with. We also annotate the pairwise relationships between the subject and the objects.

quences each. Each sequence has a high-level activity category. We annotated atomic actions in each of these videos by 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.

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.

4. Cooperative Compositional Action Understanding (CCAU)

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

---

[1]We here refer to the "task classes" in [21]

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.

4.1. 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 [46] and treat it as an image input. Following numerous other works [47, 40], 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.

4.2. 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 simul-

taneously 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 [48, 37]. 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) [49, 50, 51] 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 [27], we use an NCE loss over our feature embeddings c described in Eq. (1). c_i^m represents the feature embedding for the mth 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:

$$\mathcal{L}_{align}^{m,m'} = -\sum_i \left(\log \frac{\exp(c_i^m \cdot c_i^{m'})}{\sum_j \exp(c_i^m \cdot c_j^{m'})} \right). \tag{1}$$

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 [52] 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. (2). Where τ refers to the temperature. Further details are provided in the appendix.

$$c_{agg} = \sum_{i,j} p_{i,j} \cdot c_{i,j} \;,\; p_{i,j} = \frac{\exp(\alpha_{i,j} / \tau)}{\sum_{a,b} \exp(\alpha_{a,b} / \tau)} \tag{2}$$

4.3. 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 [53]. The numbers reported in the paper represent use the first approach with $\lambda = 10$. For details refer to the appendix.

5. 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.

5.1. 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 4.1.
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 [31] as the encoder $f(\cdot)$. Following [54], 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

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

$g(\cdot)$. 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 [55] 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 final representations, we pass this through fully connected layers to compute the final classification where we use a dropout of $p = 0.5$.
Audio. To process audio clips, we convert audio to MP3 format, compute log-mel spectrograms [46], 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.

5.2. 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.

5.2.1 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 2 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 2 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.

5.2.2 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 3 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 4 summarizes our results. The Mean Average Precision scores for each modality are reported.

5.2.3 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 summarizes the results of the cross-modality co-training experiment with and without attention module. With attention, the

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 3: 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 4: Effect of co-training encoders with different modalities on atomic action classification. The numbers reported are support weighted mAP scores.

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

Table 5: 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.

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 both activity and atomic actions without cooperation and (4) CCAU - cooperatively train with both video and atomic actions. In Table 6, we see a consistent improvement across both activity and atomic-action performance.

5.2.4 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

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 6: 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 7: 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.

models on all examples of novel classes in the validation set. **Results.** We report few-shot experiment performance in Table 7. 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 improvement.

5.3. 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 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.
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. 4 summarizes our results.
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).

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
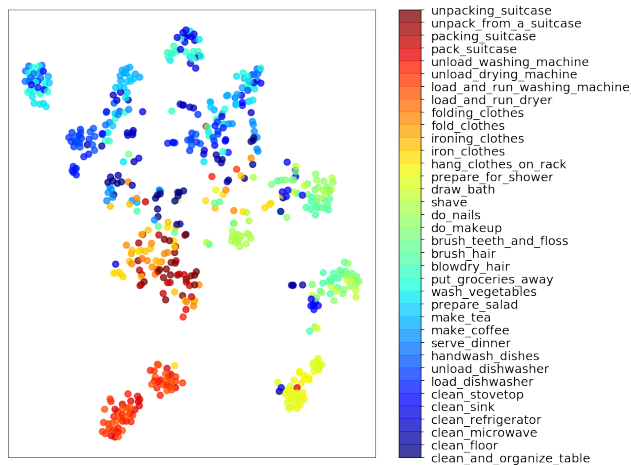

Figure 4: t-SNE visualization of Ego-View features from *CCAU* trained with ego, $3rd$ and audio modalities. The color mapping represents the relationships between the action classes, e.g., Red: Clothes; Green: Grooming; Blue: Kitchen. CCAU is able to learn meaningful clusters by utilizing compositional information.

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 spatiotemporal 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.

References

[1] F. Rezazadegan, S. Shirazi, B. Upcroft, and M. Milford, "Action recognition: From static datasets to moving robots,"

in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3185–3191, IEEE, 2017.

[2] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[3] T. Santini, W. Fuhl, and E. 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*, pp. 2594–2605, 2017.

[4] A. Haque, A. Milstein, and L. Fei-Fei, "Illuminating the dark spaces of healthcare with ambient intelligence," *Nature*, vol. 585, no. 7824, pp. 193–202, 2020.

[5] M. Lu, K. Poston, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, K. M. Pohl, J. C. Niebles, and E. 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*, pp. 637–647, Springer, 2020.

[6] A. Jalal, S. Kamal, and D. 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*, vol. 4, no. 4, 2017.

[7] Z. Luo, J.-T. Hsieh, N. Balachandar, S. Yeung, G. Pusiol, J. Luxenberg, G. Li, L.-J. Li, N. L. Downing, A. Milstein, *et al.*, "Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring," *Machine Learning for Healthcare (MLHC)*, vol. 2, 2018.

[8] U. Blanke and B. Schiele, "Remember and transfer what you have learned-recognizing composite activities based on activity spotting," in *International Symposium on Wearable Computers (ISWC) 2010*, pp. 1–8, IEEE, 2010.

[9] J. M. Zacks and B. Tversky, "Event structure in perception and conception.," *Psychological bulletin*, vol. 127, no. 1, p. 3, 2001.

[10] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10236–10247, 2020.

[11] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[12] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 510–526, Springer, 2016.

[13] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

[14] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[15] D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–736, 2018.

[16] S. Malla, B. Dariush, and C. Choi, "Titan: Future forecast using action priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2020.

[17] C.-Y. Chang, D.-A. Huang, D. Xu, E. Adeli, L. Fei-Fei, and J. C. Niebles, "Procedure planning in instructional videos," in *European Conference on Computer Vision*, pp. 334–350, Springer, 2020.

[18] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE international conference on computer vision*, pp. 2630–2640, 2019.

[19] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "Coin: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216, 2019.

[20] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI Conference on Artificial Intelligence*, pp. 7590–7598, 2018.

[21] B. Jia, Y. Chen, S. Huang, Y. Zhu, and S.-C. Zhu, "Lemma: A multi-view dataset for learning multi-agent multi-task activities," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 767–786, 2020.

[22] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion.," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[24] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[25] K. Simonyan and A. 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, (Cambridge, MA, USA), p. 568–576, MIT Press, 2014.

[26] N. Sayed, B. Brattoli, and B. Ommer, "Cross and learn: Cross-modal self-supervision," *CoRR*, vol. abs/1811.03879, 2018.

[27] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," *CoRR*, vol. abs/1906.05849, 2019.

[28] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.

[29] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[30] J. Carreira, A. Zisserman, and Q. Vadis, "Action recognition? a new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4733, 2018.

[31] K. Hara, H. Kataoka, and Y. 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*, pp. 6546–6555, 2018.

[32] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pp. 9–14, IEEE, 2010.

[33] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding," *arXiv preprint arXiv:1703.07475*, 2017.

[34] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1147–1153, IEEE, 2011.

[35] A. Shahroudy, J. Liu, T.-T. Ng, and G. 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*, pp. 1010–1019, 2016.

[36] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images.," *plan, activity, and intent recognition*, vol. 64, 2011.

[37] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8658–8667, 2019.

[38] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *European Conference on Computer Vision*, pp. 510–526, Springer, 2016.

[39] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," *CoRR*, vol. abs/1609.02612, 2016.

[40] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," *arXiv preprint arXiv:1807.00230*, 2018.

[41] Z. Luo, B. Peng, D. Huang, A. Alahi, and L. Fei-Fei, "Unsupervised learning of long-term motion dynamics for videos," *CoRR*, vol. abs/1701.01821, 2017.

[42] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. 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*, pp. 6047–6056, 2018.

[43] H. Kuehne, A. Arslan, and T. 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*, pp. 780–787, 2014.

[44] D. S. Hamermesh, H. Frazis, and J. Stewart, "Data watch: The american time use survey," *Journal of Economic Perspectives*, vol. 19, no. 1, pp. 221–232, 2005.

[45] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as compositions of spatio-temporal scene graphs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10236–10247, 2020.

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

[47] T. Afouras, A. Owens, J. S. Chung, and A. Zisserman, "Self-supervised learning of audio-visual objects from video," *arXiv preprint arXiv:2008.04237*, 2020.

[48] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[49] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 297–304, 2010.

[50] A. Mnih and K. Kavukcuoglu, "Learning word embeddings efficiently with noise-contrastive estimation," in *Advances in neural information processing systems*, pp. 2265–2273, 2013.

[51] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[52] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 631–648, 2018.

[53] A. Kendall, Y. Gal, and R. 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*, pp. 7482–7491, 2018.

[54] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Workshop on Large Scale Holistic Video Understanding, ICCV*, 2019.

[55] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors (2012)," *arXiv preprint arXiv:1207.0580*, 2012.