Pattern4Ego: Learning Egocentric Video Representation Using Cross-video Activity Patterns

Anonymous ECCV submission

Paper ID xxx

Abstract. With the development of Embodied AI, Robotics and Augmented Reality, videos captured from the 'first-person' point of view, also known as egocentric videos, are arousing interests in Computer Vision and Robotics communities. Further, learning a proper representation of egocentric videos can benefit diverse downstream tasks like action forecasting and human object interactions. However, current works mostly focus on learning the temporal or topological information for egocentric video representations, while the activity patterns, which reveal the behavior regularities or the intentions of people or robots in a more explicit way, are not carefully considered. In this paper, we propose a novel framework, Pattern4Ego, that learns the representations of egocentric videos using cross-video activity patterns. This framework achieves state-of-the-art performance on two representative egocentric video tasks: long-term action anticipation and context-based environment affordance. **Please notice that I leave out very important experiment result and detailed methodologies because of confidentiality**

1 Introduction

Unlike 'third-person' videos where cameras are posed in the hands of bystanders, 'first-person' videos, also known as egocentric videos, are captured by a wearable camera typically worn on the head or the chest of people. Egocentric videos are attracting more and more attention from researchers in recent years, with a lot of challenging tasks, such as monitoring human-object interactions [3, 8, 43], detecting gaze [35, 23], creating daily life activity summaries [39, 32, 63, 38], inferring the camera wearer's identity or body pose [56, 57, 27, 21, 26, 4, 45], and action recognition [60, 28]. However, most tasks only require the method to leverage nearby frames, and it remains challenging for more advanced tasks that require considering distant frames in egocentric videos.

Compared with third-person videos, egocentric videos are more complicated, and not well-studied. The reason is that, the understanding of egocentric video requires 3D analysis of the camera wearer's surrounding environment. However, the environments, including the background images and the foreground objects, in egocentric videos, are changing frequently, as the person is usually walking

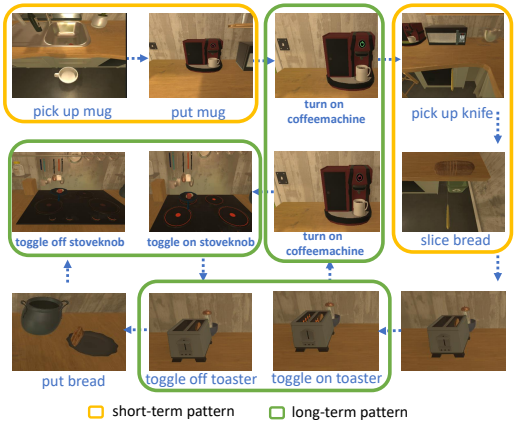


Fig. 1. Examples of long-term and short-term activity patterns. Each image represents a video clip ('visit') where a series of frames are under a similar activity. The blue arrows indicate the time order. The yellow boxes indicate the short-term pattern that are composed of two adjacent visits, while the green boxes indicate the long-term activities patterns, which can reveal the person's activity regularities or intentions but is harder to extract compared with the direct short-term patterns.

from one location to another, interacting with multiple objects and doing multiple tasks. Those changes and interactions in videos should be interpreted as high-level social behaviors based on human behavior regularities and intentions.

Many studies have tried to learn the representation of egocentric videos. Training straight-forward temporal sequential models, such as Recurrent Neural Networks (RNN), in an end-to-end data-driven manner, is an intuitive method for learning this kind of representation. However, due to the frequent change of the scene and object and the lack of training data, it is hard for an RNN to efficiently extract the representation of egocentric videos without a huge model and a huge amount of videos. Instead of this fashion, recent studies have been trying to extract high-level and sparse information from egocentric videos for a better representation. Simultaneous localization and mapping (SLAM), which models the environment via dense geometric reconstructions, suffers from SLAM failure. The current state-of-the-art method, EGO-TOPO [44], encodes a video into a low-dimensional topological graph that merges consecutive frames with similar actions into a 'visit', clusters distant visits of similarities in actions or locations into a node, and then applies Graph Convolutional Networks (GCN) [29] on the graph to aggregate the information of adjacent nodes.

However, although those methods may easily extract the relations between adjacent activities in egocentric videos, it may be very hard for them to learn the long-term relations, which could be very helpful to understand human behaviors in egocentric videos. That is because, in our daily life, in order to accomplish certain tasks, a sequence of multiple different activities are usually involved, where two distant activities often have strong relations and always exist with each other

due to regularities and intentions in people's behaviors. For example, a person may first turn on the hob, then do some cooking, and finally, turn off the hob. `turn on hob` and `turn off hob` may be distant activities with multiple and diverse actions like `pour oil`, `fry eggs`, `open fridge` in between, but they have strong relations and always exist with each other. Another example is shown in the right part of Fig.1. In order to sprinkle oregano on salad, a person needs to perform a sequence of activities as indicated by the arrow. `Sprinkle oregano` and `hang cloth` are two distant activities but with strong relationships, because a person often hangs cloth after he uses it to dry hands which are stained by seasonings like oregano. However, as shown in the figure, a person can perform multiple actions between these two activities. Besides kitchens, in other scenarios, there are more similar cases such as `running` and `take showers`, `take on shoes` and `go out`, `turn on light` and `turn off light`, etc. These kinds of activities play essential roles in facilitating the reasoning in egocentric video tasks because there exist strong correlations among these activities, and learning these activities and correlations helps to understand the high-level social regularities and intentions of human behaviors.

The above-mentioned correlations among activities in egocentric videos is yet an undiscovered zone for current methods. In this study, we propose a novel framework, Pattern4Ego, that focuses on discovering and leveraging activity patterns revealing the human intentions and behavior regularities among multiple activities in the egocentric videos. Specifically, we find that temporally non-adjacent relations or relations among rare activities are highly beneficial to improve the learning and generalization abilities of egocentric videos representations, and thus we propose a criterion and utilize statistical hypothesis testing to mine those kinds of activity patterns efficiently. Next, we build a graph through the video and the mined activity patterns to leverage the activity and pattern information we extract, and use a GCN to aggregate the information. Moreover, we employ a Gated Recurrent Units (GRU) [6] to better and especially aggregate the relations between patterned activities, and find it improves the generalization ability of the learned representation.

We evaluate our proposed framework over two datasets EPIC-Kitchens [7] and EGTEA+ [36] and on two downstream tasks in which the reasoning and prediction require the understanding of long-term correlations between activities in video. The quantitative comparisons and ablation studies demonstrate the effectiveness of our method. The learned representations can facilitate downstream tasks based on egocentric videos.

In summary, our main contributions are as follows:

- We propose that extracting and leveraging activity patterns, which reveal the relations between people's activities, help in the learning and generalization abilities of egocentric videos representations;
- We propose a novel framework, Pattern4Ego, that extracts and aggregates the activity patterns for learning the representations of egocentric videos;
- Experiments and ablation studies conducted on long-term action anticipation and long-term action recognition tasks, over EPIC-Kitchens and EGTEA+

datasets, demonstrate the superiority of the egocentric video representations learned through our proposed framework.

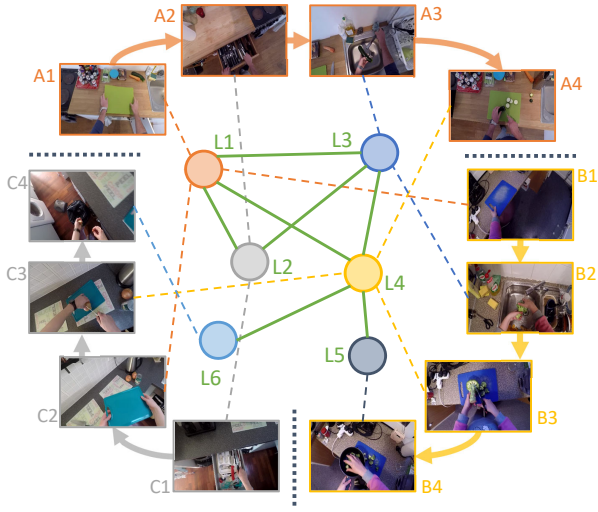
2 Related Work

Video Representation Learning is an important topic in computer vision. Current methods mainly focus on leveraging information inside a frame or among adjacent frames. There are modules to extract and aggregate action information in videos to recognize human activity [14, 24, 25, 34, 54, 33]. To explicitly reveal the relationships between objects, methods use graph to encode videos with nodes representing objects and edges to show their semantic or spatio-temporal relationships [40, 61, 1, 64]. As for feature learning in videos, to exploit temporal coherence among consecutive video frames, cycle consistency is developed [10, 49, 62, 9]. Video frame sorting is also very useful in frame prediction in video representation learning [16, 31]. Recently, more work using raw videos as input to predict appearance statistics and motion [58], encodings [18, 19, 37] and speed [2, 59]. Unlike any of those methods, our Pattern4Ego learns egocentric video representations using cross-video activity patterns.

Egocentric Videos take first-hand information from the person interacting with surroundings, where the person is usually walking around doing diverse kinds of activities, and the scene is changing frequently. With many new benchmarks appearing in recent years, the research for egocentric video is growing rapidly [15, 7, 36, 48, 53, 22, 55, 42]. Based on egocentric videos, there exist many interesting downstream task, such as hand-object interactions [3, 8, 43], gaze detection [35, 23], camera wearer's identification or body pose estimation [56, 57, 27, 21, 26, 4, 45], action recognition [60, 28] and anticipation [44]. Among those tasks, long-term prediction tasks like long-term action recognition and long-term action anticipation require the method to be able to reason through long-term information in egocentric videos such as the person's activity regularities.

Traditional solutions for such long-term prediction tasks use SLAM to achieve dense metric measurements geometrically for activity predictions [17, 50, 46], but suffer from SLAM failures Due to the complexity of egocentric videos and the large amount of frames, it is hard for a straightforward sequential model, such as RNN, to efficiently extract the long-term information in an end-to-end manner. To tackle this problem, a recent study, EGO-TOPO [44], organizes the video frames into a topological graph However, the topological map can only extract relationships between actions that are temporally adjacent. While, in practice, although some frames are far away in the time dimension, they are sequentially highly related for revealing the person's activity intentions.

Different from those above-mentioned methods, our proposed framework, Pattern4Ego, puts more focus on explicitly extracting and exploiting the activity patterns for different related activities even though they are far away in time, which can better reveal the person's activity regularities and intentions, and thus further facilitating the learning of egocentric video representations.



Pattern	Visit Sequence	Meaning
(L1, L4)	(A1, A4), (B1, B3), (C2, C3)	[place-down cutting-board, cut vegetable]
(L1, L3)	(A1, A3), (B1, B2)	[place-down cutting-board, wash vegetable]
(L2, L4)	(A2, A4), (C1, C3)	[pick-up knife, cut vegetable]
(L3, L4)	(A3, A4), (B2, B3)	[wash vegetable, cut vegetable]
(L1, L3, L4)	(A1, A3, A4), (B1, B2, B3)	[place-down cutting-board, wash vegetable, cut vegetable]

Fig. 2. Cross-video Activity Pattern Mining. For different videos with diverse environments, we cluster visits across these videos into different activity zones and extract patterns through these cross-video activity zones. **(Left)** We show 3 example videos each of which contains 4 visits, and the total 12 visits are clustered into 6 activity pattern nodes. **(Right)** We show 5 example activity patterns with their corresponding node indexes, visit indexes and meanings in the table.

3 Method

In this section, we first formulate our problem, describe the terminology and give an overview of our proposed Pattern4Ego framework. Afterwards, we step by step describe how our framework extracts and leverages cross-video activity patterns for learning egocentric video representations.

3.1 Problem Formulation, Terminology and Method Overview

Problem Formulation. Given an egocentric video V that is composed of a sequence of T frames ($V = \{f_1, ..., f_T\}$), our framework, Pattern4Ego, encodes this video V into its representation, and further uses the representation to complete downstream tasks that require long-term reasoning capabilities.

Terminology. EGO-TOPO [44] introduces the concept of 'visit' v, and our framework uses this terminology as a unit of a sequence of frames to build the graph for egocentric video representations. As proposed in EGO-TOPO, the frames set F of an egocentric video V can be segmented into multiple visits v, where a visit is a set of consecutive frames (*i.e.*, a video clip) that share a certain or similar activity, *e.g.*, an video $V = \{v_1, v_2, v_3, ...\}$, in which $v_1 = (f_1 \rightarrow f_8)$, $v_2 = (f_9 \rightarrow f_{17})$. See Fig. 3 for example, each image represents a visit (*i.e.* a set of consecutive frames in the video), which means a certain activity. The first image represents the visit with the activity `pick up the egg`, and the second image represents the visit with the activity `crash the egg`, etc.

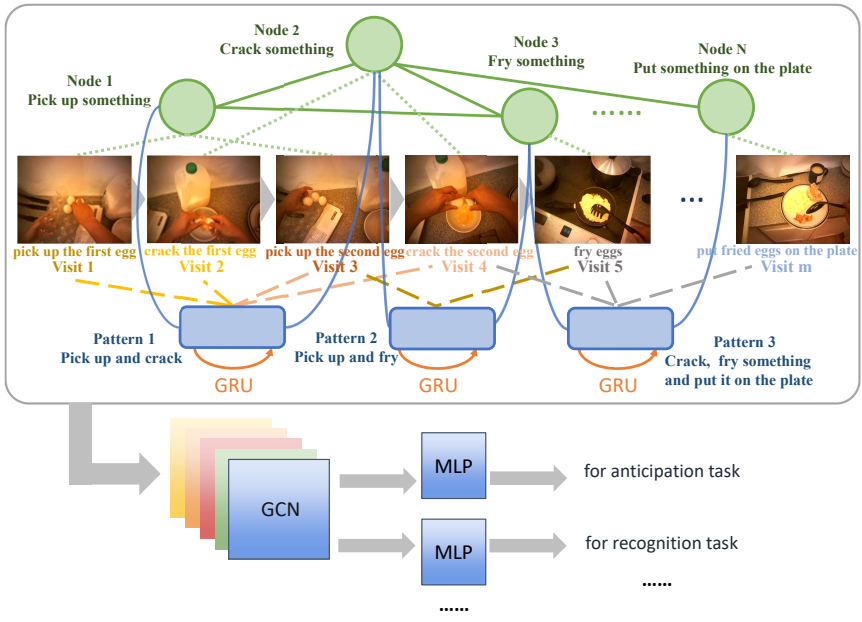


Fig. 3. Graph Construction and Information Aggregation. Our framework constructs a graph aggregating activity patterns for each video, and employs GCN through the graph to learn the representations of egocentric videos for multiple downstream tasks. In the constructed graph, an egocentric video is segmented into a sequence of visits, while nodes representing similar activities (denoted in green) and activity patterns (denoted in blue) are built on these visits. GRU (denoted in orange) are further exploited to extract the temporal information through visits with patterns.

Method Overview. Our proposed Pattern4Ego framework is composed of two main steps, **Cross-video Activity Pattern Mining** and **Graph Construction and Information Aggregation**, which are shown respectively in Fig. 2 and Fig. 3. In the **Cross-video Activity Pattern Mining** step, we ground M egocentric videos into M sequences of 'visit' representing different human activities, and then cluster these visits into several cross-video activity zones denoted as nodes. The visits in each node share a certain or similar activity. Next, we extract meaningful activity patterns using *statistical hypothesis testing* through visits with their corresponding activity zones and add new 'pattern' nodes to each graph representation of each video. In the **Graph Construction and Information Aggregation** step, we build a graph through the activity zones and activity patterns extracted in the previous step, apply GCN over the graph and thus get the representation of the video. Further, to better exploit temporal information of visits in an activity pattern, we adopt GRU through visits in patterns. At last, we add different MultiLayer Perceptrons (MLPs) after the representation for completing different downstream tasks.

More details of our framework are given in the following subsections.

3.2 Cross-video Activity Pattern Mining

The first step of our proposed framework aims to link visits across a large set of videos into activity zones (denoted as nodes), each node containing visits with a certain or similar activity. Then, we mine the activity patterns with abnormal high frequencies compared to the *i.i.d.* case that reveal people's intentions and activity regularities through linked cross-video visits.

Cross-video Visit Activity Clustering. The first sub-step aims to cluster the visits across a diverse set of videos into activity zones, and the visits in the same zone share a certain or similar activity. Because it is hard for a single action classification or person localization network to generalize in novel videos with novel environments, we utilize a Siamese network $\mathbb{L}$ that compares the activity similarity between two frames by comparing their extracted features. Here, the Siamese network may determine two frames to be similar if they share a similar action, location or visual appearance. A consecutive sequence of frames sharing the same activity are represented into a visit, and the average of similarities between the frames in two visits is computed as the similarity between them. Further, we cluster the visits into D cross-video activity zones by sequentially employing t-SNE [41] embedding and K-Means clustering using the calculated similarity matrix across the videos in the diverse environments.

Sequential Activity Pattern Mining. In this sub-step, we aim to extract the activity patterns through the video set, in which each video is composed of a sequence of visits with activity labeled from the previous sub-step.

It is important that meaningful patterns are not merely high-frequent sub-activity-sequences. For example, `(cut potato, open fridge)` may occur much more times than `(sprinkle oregano, dry hands, hang cloth)` in the video set. However, the reason that the former sub-sequence is often seen may be that both 'potato' and 'fridge' frequently occur, while there's no essential correlation between those 2 activities.

On the other hand, `(sprinkle oregano, dry hands, hang cloth)` is rarely seen, but it's just because 2 in those 3 components (`sprinkle oregano` and `hang cloth`) are rare, while the correlation among those 3 is highly strong, as people often wash and dry hands and hang cloth after sprinkle oregano.

These examples show that, in order to mine meaningful activity patterns, we should not only consider the emerging times of a sub-activity-sequence, but also take into account the occurrence frequency of its components. That's to say, it's better to mine activity patterns with abnormally high occurrence frequencies compared with what they "should" be according to an *i.i.d.* hypothesis, *e.g.* `(sprinkle oregano, dry hands, hang cloth)`, instead of selecting activity patterns just because some of their components frequently occur, *e.g.* `(cut potato, open fridge)`.

To this end, we formulate the activity pattern mining problem to be a statistical hypothesis testing one, where the actual probability is compared with the hypothetical (*i.i.d.*) case, as a result, the activity patterns with abnormally high frequencies would be extracted.

Specifically, when an $r-$ary pattern $(d_1, \cdots, d_r), d_j \in \{1, 2, \cdots, D\}, \forall 1 \leq j \leq r$ is examined, we take the *null hypothesis* H_0 to be "components of the pattern $(d_1, \cdots, d_r)$ are *i.i.d.* in those M sequences", and the *alternative hypothesis* H_1 to be "components of the pattern $(d_1, \cdots, d_r)$ have sequential correlation within a certain sliding window length w in those M sequences".

A test statistic T is proposed to check whether those components of a pattern are *i.i.d.*: the occurrences of a sub-sequence $(d_1, \cdots, d_r)$ in a total of M sequences $s_1, \cdots, s_M$, divided by the product of occurrences of every single element in the pattern for normalization.

$$T = \frac{\sum_{i=1}^{M} \sum_{j_1 < \cdots < j_r, j_r - j_1 < w} \text{match}((s_i^{(j_1)}, \cdots, s_i^{(j_r)}), (d_1, \cdots, d_r))}{\prod_{k=1}^{m} \sum_{i=1}^{M} \sum_{j=1}^{l_i} \text{match}(s_i^{(j)}, d_k)} \tag{1}$$

$$\text{where} \qquad \text{match}(A, B) = \begin{cases} 1, A = B \\ 0, \text{otherwise} \end{cases} \tag{2}$$

where w is the size of the sliding window — sequences whose start and end are too distant, *i.e.* $j_r - j_1 > w$ in equation 1 would not be taken into account for saving computation cost.

Theorem 1. *The test statistic T is a scaled estimator of the ratio of the probability of any sub-sequence matching the pattern over such probability in* i.i.d. *case.*

$$\lim_{M \to \infty} \frac{(\sum_{i=1}^{M} l_i)^r}{\sum_{i=1}^{M} (l_i - w + 1)\binom{r}{w}} \cdot \mathbb{E}(T) = P((d_1, \cdots, d_r)) \tag{3}$$

$$\text{where} \quad P((d_1, \cdots, d_r)) = \frac{\mathbb{P}(\text{match}((s_i^{(j_1)}, \cdots, s_i^{(j_r)}), (d_1, \cdots, d_r)) = 1)}{\mathbb{P}(\text{match}((s_i^{(j_1)}, \cdots, s_i^{(j_r)}), (d_1, \cdots, d_r)) = 1 | H_0)} \tag{4}$$

Equation 3 and 4 shows that patterns with the highest T statistic have the most abnormally high probability compared to the *i.i.d.* case.

The p-value could be defined as

$$p = \mathbb{P}(T > T_0), \text{where} \quad T_0 \geq E(T). \tag{5}$$

Equation 5 shows that p is monotonically decreasing for all sufficiently large T_0. As a result, we reach the following theorem:

Theorem 2. *Patterns with the highest T statistic have the most significant p-values.*

All patterns with a statistically significant p-value can be assumed to be meaningful, so it is reasonable to formulate our problem as performing hypothesis testing on T, and then extract patterns with the highest T as the most meaningful ones for the next sub-step.

Besides, we also filtered out those patterns whose occurrences are plainly too few. This guarantees all obtained patterns are seen by a reasonable frequency and sequentially correlated, rather than selecting activity patterns just because some of their components frequently occur.

In practice, we employ the Prefix Span [51] method to boost the mining of meaningful activity patterns defined and described above.

As is mentioned before, it may be hard for RNN to efficiently encode an egocentric video, as the number of frames is large and the scene is frequently changing. However, in activity pattern nodes, the number of frames is relatively small and the sequence of visits have much stronger relations, thus making it easier for RNN to efficiently extract temporal information. Therefore, on each pattern node, we apply GRU, an outstanding recurrent neural network for handling applications involving sequential or temporal data, to better extract the temporal and relation information of visits within activity patterns. Experimental results show that GRU helps improve the performance of learned representations, especially the generalization capability in the RARE action class. Please see our experiment section for more analysis.

For different downstream tasks based on egocentric videos, we add different MLPs following the learned representations. Please see Section 4 for the downstream applications, experiments and analysis.

4 Experiments

We perform experiments and evaluate our proposed Pattern4Ego framework on two long-term reasoning tasks: long-term action anticipation task and long-term action recognition task, over two datasets: EPIC-Kitchens [7] and EGTEA+ [36], and set up many baselines for comparisons. Quantitative and qualitative results, as well as ablation studies, demonstrate the effectiveness and superiority of the proposed framework, proving that activity patterns we propose and extract facilitate the representation learning of egocentric videos.

4.1 Datasets and Tasks

Datasets. We adopt two egocentric video datasets for evaluation:

- **EGTEA+** [36] has 32 subjects and 7 recipes in a kitchen, including 53 objects and 19 kinds of actions. A video contains frames from preparing to completing a dish, with clips annotation for different action interactions, such as open door, open fridge, take tofu, wash carrot and etc.
- **EPIC-Kitchens** [7] is also a video dataset of cooking activities, containing 352 different objects and 125 different actions. Compared with EGTEA+, this dataset is larger and collected across multiple kitchens.
- **AI2THOR-EGO** is a dataset that we collect over the AI2THOR [30], an interactive 3D indoor environment that provides accurate modeling of physical world. We use a single agent to imitate human activities, and take egocentric frames to form videos. Our dataset contains 300 different videos on 120 different scenes, including kitchen, bathroom, livingroom and bedroom. Instead of mannual collection, We use a hand-crafted policy to create interactions, take frames, and generate labels. To enhance variety, we use topological sort to get action sequence, and random seeds to create different initial settings of the scene.

Tasks. We conduct the following two downstream tasks to demonstrate the effectiveness of the learned representation.

- **Long-term Action Anticipation.** This task requires the method to predict all the future actions according to the given first 25%, 50% and 75% fraction of the input video. Different from the action prediction tasks [11, 65, 12, 7, 47, 13, 52] that only predict the next adjacent action, long-term action anticipation aims to predict all future actions, which requires the method to not only have the ability to understand what is going on, but also have a global vision of current videos to further analyze the progress of current tasks and what should be done in the future to accomplish those tasks.
- **Long-term Action Recognition** We shall leave out further details because of confidentiality.

4.2 Baselines and Evaluation Metrics

Baselines. In our experiments, we compare our proposed framework with many baseline methods, in which **EGO-TOPO** [44] is the current state-of-the-art and the most important baseline. It organizes egocentric videos into a topological graph, and then adopts GCN to extract the representation of the graph. Also, we compare our framework against the following methods:

- **TRAINDIST** calculates the distribution of actions performed in all training videos, predict the actions of test videos, test if dominant actions are repeated, regardless of the content of the test video.
- **I3D** [5] samples 64 clips and averages their features as the video feature.
- **RNN** models temporal dynamics in videos using LSTM [20] layers.
- **ACTIONVLAD** [14] models temporal dynamics with non-uniform pooling.
- **VIDEOGRAPH** [25] and **TIMECEPTION** [24] build complex temporal models using multi-scale temporal convolutions or attention mechanisms over learned latent concepts from clip features over large time scales.

Evaluation Metric. For both tasks on both datasets, we exploit mean average precision (mAP) of multi-label classification as the evaluation metric. Using this metric, we evaluate our method over three kinds of action classes: the class of all actions (denoted as the all setting ALL), the class of actions with fewer than 10 instances (denoted as the rare setting $RARE$) and the class of action with more than 100 instances (denoted as the frequent setting $FREQ$).

4.3 Quantitative Results and Analysis

Please see tables above, we will leave out further information because of confidentiality.

4.4 Ablation Studies and Analysis

To further demonstrate the necessity of the different components of our framework, we conduct ablation studies by comparing our method with:

1) Our method without GRU (Ours w/o GRU);

Table 1. Quantitative comparisons of long-term action anticipation task. Our method outperforms all the baselines by a large margin on all datasets and metrics. And we will leave out some data because of confidentiality.

Dataset	**EPIC-Kitchens**			**EGTEA+**		
mAP	ALL	FREQ	RARE	ALL	FREQ	RARE
TRAINDIST	16.5	39.1	5.7	59.1	68.2	35.2
I3D	32.7	53.3	23.0	72.1	79.3	53.3
RNN	32.6	52.3	23.3	70.4	76.6	54.3
ACTIONVLAD	29.8	53.5	18.6	73.3	79.0	58.6
VIDEOGRAPH	22.5	49.4	14.0	67.7	77.1	47.2
TIMECEPTION	35.6	55.9	26.1	74.1	79.7	59.7
EGO-TOPO	38.0	56.9	29.2	73.5	80.7	54.7
Ours (binary)	xx	xx	xx	xx	xx	xx
Ours (ternary)	xx	xx	xx	xx	xx	xx

2) Our method without GCN (Ours w/o GCN).

We conduct ablation experiments on both tasks over both datasets, and the results on two tasks are shown in Table 3 and Table 4. The results clearly show that both GCN and GRU help in improving the performance of our framework.

Specifically, GRU plays an essential role in improving the performance on *RARE* actions, demonstrating that extracting the temporal information of activity patterns equips the representation with stronger reasoning capabilities. Note that the RNN baseline in Table 1 and 2 perform much worse than our framework, while both of them employ a recurrent neural network. The reason is that the visits in the activity patterns have strong correlations, making it much easier for the RNN to extract meaningful information.

4.5 Qualitative Results and Analysis

Fig. 4 illustrates the example results and patterns extracted by our method on the long-term action anticipation task. The activity patterns could significantly help the anticipation task. In the third row, the extracted activity pattern `(open drawer, pick up knife, pick up pad)` could anticipate another action of opening the drawer in order to put back the tools; in the fourth row, the extracted activity pattern `(pick up towel, dry container)` means that the person doesn't need to dry the container again in the latter part of the video.

5 Conclusion

In this paper, we study the problem of learning the egocentric video representations. Our proposed framework, Pattern4Ego, is the first that explicitly extracts and leverages the activity patterns revealing the regularities and intentions of

Table 2. Quantitative comparisons of long-term action recognition task. Our proposed method outperforms most baselines, achieves comparable performance in the ALL metric with TIMECEPTION baseline, and outperforms it in the *RARE* metric.

Dataset	**EPIC-Kitchens**			**EGTEA+**		
mAP	ALL	FREQ	RARE	ALL	FREQ	RARE
I3D	38.1	64.2	25.7	74.7	85.2	52.0
RNN	36.6	66.0	22.7	76.4	86.0	55.5
TIMECEPTION	**45.4**	**72.0**	32.8	75.8	**88.1**	49.2
EGO-TOPO	40.7	64.4	29.7	71.7	79.2	50.3
Ours (binary)	xx	xx	xx	**xx**	84.3	**61.4**
Ours (ternary)	xx	xx	**34.2**	xx	xx	xx

Table 3. Ablation study on long-term action anticipation task. It is clear GCN and GRU help to improve the performance of our framework. GRU helps a lot in improving the performance in the *RARE* metric.

Dataset	**EPIC-Kitchens**			**EGTEA+**		
mAP	ALL	FREQ	RARE	ALL	FREQ	RARE
Ours w/o GRU	40.7	xx	32.0	xx	xx	xx
Ours w/o GCN	xx	xx	xx	xx	xx	xx
Ours	**41.8**	**xx**	**33.5**	**xx**	**xx**	**xx**

human behaviors for better egocentric video representations. The quantitative and qualitative evaluations, ablation studies and analysis, show the effectiveness of our framework, and the importance of different components of our method.

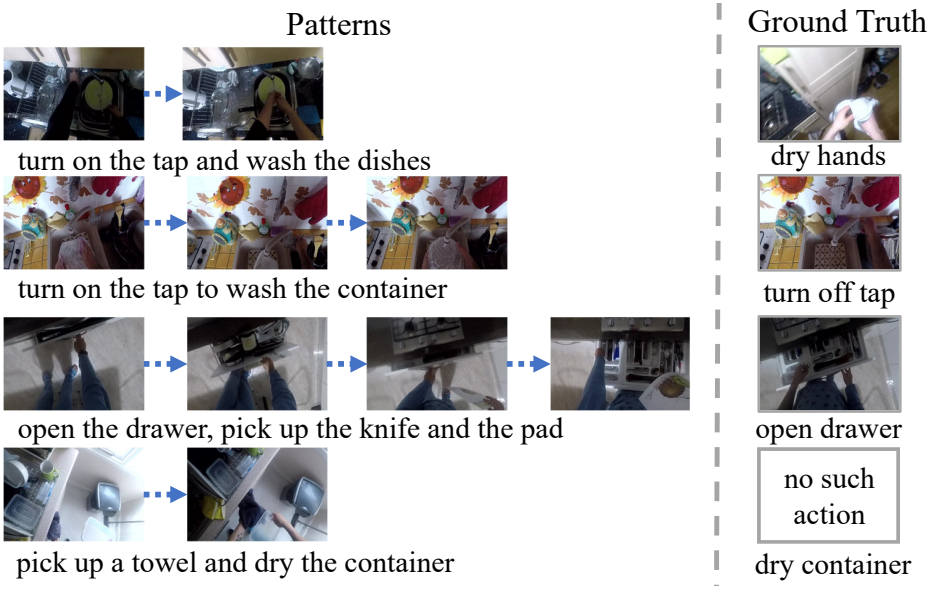


Fig. 4. Example results of long-term action anticipation task. The left columns show patterns extracted in the input video, and the rightmost column shows one of the ground truth actions. The empty frame denotes the action not happening.

Table 4. Ablation study on long-term action recognition task. It is clear that GCN and GRU help improve the performance of our framework. GRU helps a lot in improving the performance in the *RARE* metric.

Dataset	**EPIC-Kitchens**			**EGTEA+**		
mAP	ALL	FREQ	RARE	ALL	FREQ	RARE
Ours w/o GRU	xx	xx	xx	74.6	xx	xx
Ours w/o GCN	xx	65.4	xx	73.6	**xx**	xx
Ours	**xx**	**xx**	**xx**	**xx**	xx	**xx**

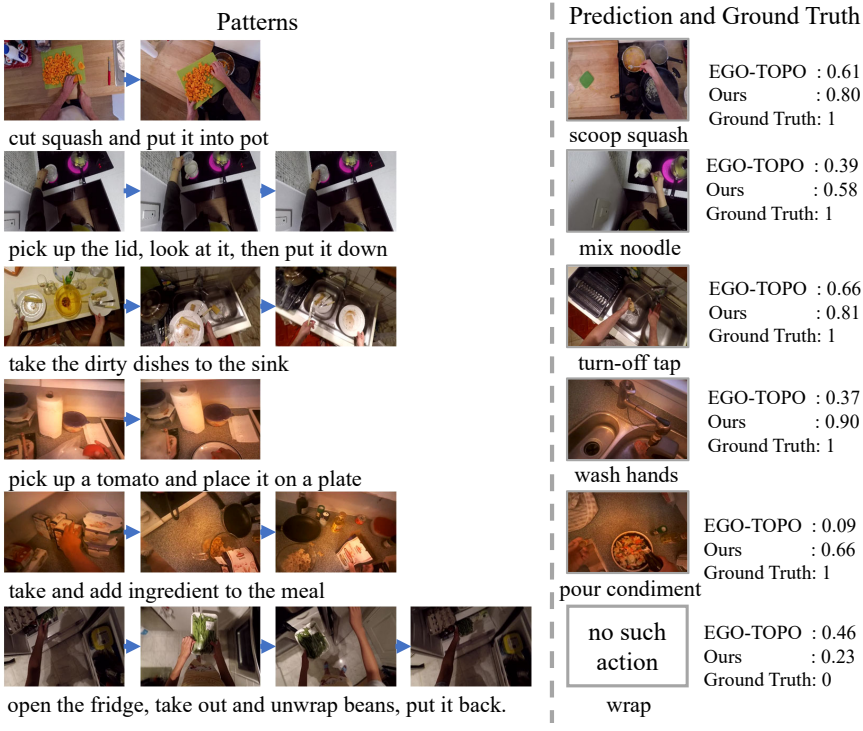


Fig. 5. Qualitative comparisons on long-term anticipation task between our method and EGO-TOPO. The left columns show our extracted patterns from video inputs, and the rightmost column shows one ground truth action and the predictions of our method and EGO-TOPO. The empty frame denotes the action not happening.

References

1. Baradel, F., Neverova, N., Wolf, C., Mille, J., Mori, G.: Object level visual reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 105–121 (2018)
2. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: Learning the speediness in videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9922–9931 (2020)
3. Cai, M., Kitani, K.M., Sato, Y.: Understanding hand-object manipulation with grasp types and object attributes. In: Robotics: Science and Systems. vol. 3. Ann Arbor, Michigan; (2016)
4. Cai, M., Lu, F., Sato, Y.: Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
6. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
7. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 720–736 (2018)
8. Damen, D., Leelasawassuk, T., Mayol-Cuevas, W.: You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. Computer Vision and Image Understanding **149**, 98–112 (2016)
9. Dong, J., Shuai, Q., Zhang, Y., Liu, X., Zhou, X., Bao, H.: Motion capture from internet videos. In: European Conference on Computer Vision. pp. 210–227. Springer (2020)
10. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1801–1810 (2019)
11. Furnari, A., Battiato, S., Grauman, K., Farinella, G.M.: Next-active-object prediction from egocentric videos. Journal of Visual Communication and Image Representation **49**, 401–411 (2017)
12. Furnari, A., Farinella, G.M.: What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6252–6261 (2019)
13. Gao, J., Yang, Z., Nevatia, R.: Red: Reinforced encoder-decoder networks for action anticipation. arXiv preprint arXiv:1707.04818 (2017)
14. Girdhar, R., Ramanan, D., Gupta, A., Sivic, J., Russell, B.: Actionvlad: Learning spatio-temporal aggregation for action classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 971–980 (2017)
15. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., Martin, M., Nagarajan, T., Radosavovic, I., Ramakrishnan, S.K., Ryan, F., Sharma, J., Wray, M., Xu, M., Xu, E.Z., Zhao, C., Bansal, S., Batra, D., Cartillier, V., Crane, S., Do, T., Doulaty, M., Erapalli, A., Feichtenhofer, C., Fragomeni, A., Fu, Q., Fuegen, C., Gebreselasie, A., Gonzalez,

C., Hillis, J., Huang, X., Huang, Y., Jia, W., Khoo, W., Kolar, J., Kottur, S., Kumar, A., Landini, F., Li, C., Li, Y., Li, Z., Mangalam, K., Modhugu, R., Munro, J., Murrell, T., Nishiyasu, T., Price, W., Puentes, P.R., Ramazanova, M., Sari, L., Somasundaram, K., Southerland, A., Sugano, Y., Tao, R., Vo, M., Wang, Y., Wu, X., Yagi, T., Zhu, Y., Arbelaez, P., Crandall, D., Damen, D., Farinella, G.M., Ghanem, B., Ithapu, V.K., Jawahar, C.V., Joo, H., Kitani, K., Li, H., Newcombe, R., Oliva, A., Park, H.S., Rehg, J.M., Sato, Y., Shi, J., Shou, M.Z., Torralba, A., Torresani, L., Yan, M., Malik, J.: Ego4d: Around the World in 3,000 Hours of Egocentric Video. CoRR **abs/2110.07058** (2021), https://arxiv.org/abs/2110.07058

16. Griffin, B.A., Corso, J.J.: Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8914–8923 (2019)
17. Guan, J., Yuan, Y., Kitani, K.M., Rhinehart, N.: Generative hybrid representations for activity forecasting with no-regret learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 173–182 (2020)
18. Han, T., Xie, W., Zisserman, A.: Video representation learning by dense predictive coding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
19. Han, T., Xie, W., Zisserman, A.: Memory-augmented dense predictive coding for video representation learning. In: European conference on computer vision. pp. 312–329. Springer (2020)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
21. Hoshen, Y., Peleg, S.: An egocentric look at video photographer identity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4284–4292 (2016)
22. Hu, T., Sarkar, K., Liu, L., Zwicker, M., Theobalt, C.: Egorenderer: Rendering human avatars from egocentric camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14528–14538 (2021)
23. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 754–769 (2018)
24. Hussein, N., Gavves, E., Smeulders, A.W.: Timeception for complex action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 254–263 (2019)
25. Hussein, N., Gavves, E., Smeulders, A.W.: Videograph: Recognizing minutes-long human activities in videos. arXiv preprint arXiv:1905.05143 (2019)
26. Jiang, H., Grauman, K.: Seeing invisible poses: Estimating 3d body pose from egocentric video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3501–3509. IEEE (2017)
27. Jiang, H., Ithapu, V.K.: Egocentric pose estimation from human vision span. arXiv preprint arXiv:2104.05167 (2021)
28. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
29. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
30. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: AI2-THOR: An Interactive 3D Environment for Visual AI. arXiv (2017)

31. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE international conference on computer vision. pp. 667–676 (2017)
32. Lee, Y.J., Grauman, K.: Predicting important objects for egocentric video summarization. International Journal of Computer Vision **114**(1), 38–55 (2015)
33. Li, X., Liu, C., Shuai, B., Zhu, Y., Chen, H., Tighe, J.: Nuta: Non-uniform temporal aggregation for action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3683–3692 (2022)
34. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 909–918 (2020)
35. Li, Y., Fathi, A., Rehg, J.M.: Learning to predict gaze in egocentric video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3216–3223 (2013)
36. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of the European conference on computer vision (ECCV). pp. 619–635 (2018)
37. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104 (2016)
38. Lu, C., Liao, R., Jia, J.: Personal object discovery in first-person videos. IEEE Transactions on Image Processing **24**(12), 5789–5799 (2015)
39. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2714–2721 (2013)
40. Ma, C.Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., Graf, H.P.: Attend and interact: Higher-order object interactions for video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6790–6800 (2018)
41. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
42. Makansi, O., Cicek, O., Buchicchio, K., Brox, T.: Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
43. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8688–8697 (2019)
44. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: CVPR (2020)
45. Ng, E., Xiang, D., Joo, H., Grauman, K.: You2me: Inferring body pose in egocentric video via first and second person interactions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
46. Park, H.S., Hwang, J.J., Niu, Y., Shi, J.: Egocentric future localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4697–4705 (2016)
47. Pirri, F., Mauro, L., Alati, E., Ntouskos, V., Izadpanahkakhk, M., Omrani, E.: Anticipation and next action forecasting in video: an end-to-end model with memory. arXiv preprint arXiv:1901.03728 (2019)
48. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2847–2854. IEEE (2012)

49. Reda, F.A., Sun, D., Dundar, A., Shoeybi, M., Liu, G., Shih, K.J., Tao, A., Kautz, J., Catanzaro, B.: Unsupervised video interpolation using cycle consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 892–900 (2019)
50. Rhinehart, N., Kitani, K.M.: First-person activity forecasting with online inverse reinforcement learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3696–3705 (2017)
51. Sharma, P., Balakrishna, G.: Prefixspan: Mining sequential patterns by prefix-projected pattern. International Journal of Computer Science and Engineering Survey **2**(4), 111 (2011)
52. Shi, Y., Fernando, B., Hartley, R.: Action anticipation with rbf kernelized feature mapping rnn. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 301–317 (2018)
53. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018)
54. Tang, J., Xia, J., Mu, X., Pang, B., Lu, C.: Asynchronous interaction aggregation for action detection. In: European Conference on Computer Vision. pp. 71–87. Springer (2020)
55. Thapar, D., Nigam, A., Arora, C.: Anonymizing egocentric videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2320–2329 (2021)
56. Tome, D., Peluse, P., Agapito, L., Badino, H.: xr-egopose: Egocentric 3d human pose from an hmd camera. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7728–7738 (2019)
57. Wang, J., Liu, L., Xu, W., Sarkar, K., Theobalt, C.: Estimating egocentric 3d human pose in global space. arXiv preprint arXiv:2104.13454 (2021)
58. Wang, J., Jiao, J., Bao, L., He, S., Liu, Y., Liu, W.: Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4006–4015 (2019)
59. Wang, J., Jiao, J., Liu, Y.H.: Self-supervised video representation learning by pace prediction. In: European conference on computer vision. pp. 504–521. Springer (2020)
60. Wang, X., Zhu, L., Wang, H., Yang, Y.: Interactive prototype learning for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8168–8177 (2021)
61. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Proceedings of the European conference on computer vision (ECCV). pp. 399–417 (2018)
62. Wu, H., Wang, X.: Contrastive learning of image representations with cross-video cycle-consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10149–10159 (2021)
63. Yonetani, R., Kitani, K.M., Sato, Y.: Visual motif discovery via first-person vision. In: European Conference on Computer Vision. pp. 187–203. Springer (2016)
64. Zhang, Y., Tokmakov, P., Hebert, M., Schmid, C.: A structured model for action detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9975–9984 (2019)
65. Zhou, Y., Berg, T.L.: Temporal perception and prediction in ego-centric video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4498–4506 (2015)