A Hybrid Egocentric Activity Anticipation Framework via Memory-Augmented Recurrent and One-shot Representation Forecasting

Tianshan Liu and Kin-Man Lam
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
tianshan.liu@connect.polyu.hk, enkmlam@polyu.edu.hk

Abstract

Egocentric activity anticipation involves identifying the interacted objects and target action patterns in the near future. A standard activity anticipation paradigm is recurrently forecasting future representations to compensate the missing activity semantics of the unobserved sequence. However, the limitations of current recursive prediction models arise from two aspects: (i) The vanilla recurrent units are prone to accumulated errors in relatively long periods of anticipation. (ii) The anticipated representations may be insufficient to reflect the desired semantics of the target activity, due to lack of contextual clues. To address these issues, we propose "HRO", a hybrid framework that integrates both the memory-augmented recurrent and one-shot representation forecasting strategies. Specifically, to solve the limitation (i), we introduce a memory-augmented contrastive learning paradigm to regulate the process of the recurrent representation forecasting. Since the external memory bank maintains long-term prototypical activity semantics, it can guarantee that the anticipated representations are reconstructed from the discriminative activity prototypes. To further guide the learning of the memory bank, two auxiliary loss functions are designed, based on the diversity and sparsity mechanisms, respectively. Furthermore, to resolve the limitation (ii), a one-shot transferring paradigm is proposed to enrich the forecasted representations, by distilling the holistic activity semantics after the target anticipation moment, in the offline training. Extensive experimental results on two large-scale data sets validate the effectiveness of our proposed HRO method.

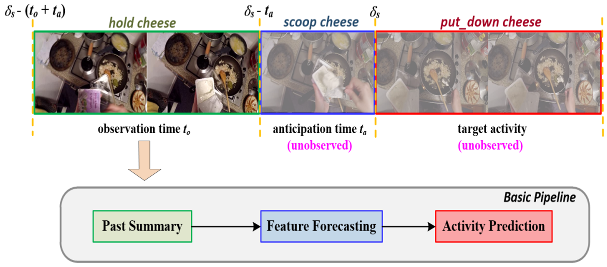


Figure 1. Illustration of task definition and basic pipeline for egocentric activity anticipation.

1. Introduction

With the popularity of wearable cameras, e.g., GoPro, egocentric perception has attracted extensive attention over the past decade [40, 51]. Among the diverse egocentric vision tasks [3, 28, 34, 45], anticipating the near future activities is a crucial high-level task, due to its wide-spread real-world applications [16, 32], e.g., human-robot interaction, autonomous driving, abnormal event alerts, etc. However, anticipating an unseen egocentric-activity before it starts is non-trivial, because of a number of challenges, such as the semantics gap between past and future events, the scarcity of useful clues in incomplete observations, and frequent ego-motion and cluttered backgrounds in egocentric videos. Therefore, egocentric activity anticipation (EAA) is still a challenging task.

As illustrated in Fig. 1, following the definition in [8], the "observation time" t_o is the temporal length of the observed video clip, and the "anticipation time" t_a indicates the temporal interval before the target anticipation moment. Thus, the goal of the EAA task is to anticipate an egocentric activity occurring at moment δ_s, by merely observing a video clip in the range of $[\delta_s - (t_o + t_a), \delta_s - t_a]$, i.e., preceding the target activity beginning at moment δ_s by a duration t_a. Recurrent forecasting [12, 54] is a commonly used paradigm for anticipating future activities, as it is flexible to predict results at any moment. Typically, the recursive anticipation model summarizes the past observations, and then forecasts what will occur in the near future. This process is continuously conducted until the target anticipation moment δ_s arrives. Therefore, the performance of a recursive anticipation system depends on whether it is able to forecast discriminative representations, which cover the activity semantics in the unseen video sequence. However, the current recursive-model-based EAA methods still struggle to achieve satisfactory performance.

We argue that the improvement is impeded by two major obstacles. First, the vanilla recursive model, based on recurrent neural networks (RNNs), is prone to accumulated prediction errors, especially in relatively long-period anticipation. Although some methods [40, 54] attempted to introduce contrastive learning [6, 22, 38] to regulate the representation forecasting, the performance still obviously degrades when the anticipation interval increases. The underlying reason is that the RNN-based model mainly updates the memory cell states by remembering information from the previous step, which makes it hard to maintain long-term dependencies among all the past steps. In addition, given a short-length observation input with limited dynamics, it is difficult for the RNN-based model to anticipate the subsequent activities, as the cell states can only reveal the relations within the current observed short sequence, without any access to external knowledge. Second, the existing methods merely force the forecasted representations to compensate the semantics of the anticipation time, which are likely insufficient to represent the future activity. In other words, the lack of contextual cues after the target anticipation time-step poses great challenges to the anticipation model to make correct prediction, especially when the anticipation moment is at the transition boundary of two consecutive activities. For example, as shown in Fig. 1, even if the forecasted representations contain the semantics of "scoop cheese" covering the anticipation time, directly utilizing these features to infer the subsequent activity, i.e., "put_down cheese" is still ambiguous. The relations between the representations before and after the anticipation time step have not been well studied.

To address the aforementioned issues, in this paper, we propose a **h**ybrid framework in a combination of memory-augmented **r**ecurrent and **o**ne-shot representation forecasting, termed as "HRO", for egocentric activity anticipation. First, to mitigate the error accumulation issues in conventional recursive models, we introduce a memory bank into the process of recurrent representation forecasting. Different from the internal memory cells of the RNN units, the memory bank externally stores long-term prototypical activity semantics learned from training data, which are not limited to the current observation inputs. Our model is trained to reconstruct future representations by a convex combination of the memory items, via an attention-based memory addressing mechanism. To further guide the learning of the memory bank, we design two loss functions, based on a diversity scheme and a sparsity scheme, respectively, which can force the memory bank to be equipped with the desired properties. Second, to maximally incorporate contextual clues into the forecasted representation, we propose a one-shot transferring strategy to explicitly explore semantics relations between the features before and after the anticipation time-step. Specifically, in the offline training, at each target anticipation moment, we adopt a transition layer to project the features anticipated by the recurrent model, into another space to simulate the activity semantics extracted from a future video clip. This holistic transferring process is supervised by a similarity learning loss, which minimizes the semantics gap between simulated and future features.

The main contributions of this paper can be summarized in three ways. 1) We propose a memory-augmented recurrent representation forecasting paradigm, which aims to guarantee that the anticipated representations always contain the discriminative activity semantics, with the help of a compressed memory bank. Moreover, two regularization loss terms, based on the diversity and sparsity mechanisms, are designed to guide the updating of the memory bank. 2) A one-shot transferring strategy is presented to further recalibrate the forecasted representations, by injecting future activity semantics, at the target activity anticipation time step. 3) Extensive experimental results on two challenging data sets, i.e., EGTEA Gaze+ [32] and EPIC-Kitchens [9], highlight the performance improvements of our proposed hybrid framework over other state-of-the-art methods.

2. Related Work

Egocentric Activity Recognition. With the development of deep-learning-based video recognition methods and the emergence of large-scale egocentric video data sets [9, 43], a great improvement has been witnessed in egocentric activity recognition [13, 33, 50]. In the early stage, a popular pipeline was used to locate the regions involving human-object interaction by utilizing diverse frame-level annotations, such as gaze cues [25, 42], hand segmentation [35]. However, these required fine-grained annotations may be inaccessible in practice. To alleviate this issue, Ego-RNN [46] and LSTA [44] incorporated spatial attention mechanism into ConvLSTM blocks to localize relevant regions across frames, in a weakly-supervised manner. SAP [50] leveraged a detector to generate local object-centric features, which were employed as the guidance information to identify human-object interactions via a symbiotic attention module. Our work builds on the basic concept explored in egocentric activity recognition, such as the extraction of spatio-temporal representations, the exploration of multiple modalities, etc. However, different from the aforementioned works, we address the task of egocentric activity anticipation, which focuses on compensating the future semantics of the unseen sequence, rather than merely extracting discriminative features from the observed video clips.

Early Activity Recognition. The goal of early activity recognition [1, 4, 24, 39] is to predict the category of an ongoing activity, based on partial observations with incomplete executions [10]. Kong *et al*. [30] proposed a deep sequential context model to enrich the representations extracted from partially observed videos, for activity prediction. Wang *et al*. [49] employed a teacher-student learning framework, to transfer knowledge gained from an activity recognition model to the target early activity prediction model. Considering the similarity between the tasks of early recognition and anticipation, both need to forecast the representations or activity semantics in the unseen video parts. For these two online tasks, the RNN-based units, e.g., LSTMs or GRUs, have been widely utilized as the basic architecture to process streaming video. However, in the task of activity anticipation, the model is required to recognize the activity ahead before it begins, which means that the target activity cannot even be partially observed in the anticipation time.

Egocentric Activity Anticipation. Starting with the competition proposed in [8], in the past few years, various methods and frameworks [14, 15, 41] have been investigated on activity anticipation in egocentric videos. Girdhar *et al*. [18] devised an attention-based video modeling architecture, i.e., Anticipative Video Transformer (AVT), for egocentric activity anticipation. Ke *et al*. [29] proposed a time-conditioned prediction framework, by explicitly modeling the anticipation time interval as a parameter. RU-LSTM [16] leveraged multimodal data, including RGB, optical flow and object features, as inputs, and employed both rolling and unrolling LSTM blocks to improve the anticipation performance. To further tackle the error-accumulation issues of vanilla RNN-based anticipation models, SRL [40] and LAI [54] introduced contrastive learning to regularize the forecasting of future representations. Fernando et al. *et al*. [14] designed a series of Jaccard similarity measures to build the relations between past observations and future sequences. However, different from the above-mentioned works, we improve the recursive representation forecasting framework by introducing a memory bank, which can maintain prototypical activity semantics learned from the training data. Moreover, we build a hybrid anticipation framework by incorporating a one-shot transferring paradigm, to further recalibrate the forecasted representations by exploring future semantics, at the target anticipation moment.

Memory Networks. In the deep-learning domain, the memory network can be categorized into two branches. One is the internal memory, which implicitly updates in a recurrent fashion, e.g., LSTM [23], GRU [7]. The other is the external memory [19, 21, 31, 53], which can be read or written, via an attention-based scheme. According to the characteristics of the stored data, the external memory can be categorized into episodic and non-episodic types. The episodic memory [20, 52] is limited to the current observations, while the non-episodic methods [27,37] aim to learn a persistent memory. In our work, we choose the RNN-based recursive model to forecast the representations in the anticipation time interval, as the internal memory is flexible to give the anticipation results at any time-step, and is well-suited for online tasks. In addition, we also maintain an external memory bank, which can augment the recursive anticipation model to reconstruct future representations from the discriminative activity prototypes.

3. Methodology

3.1. Problem Formulation

The objective of egocentric activity anticipation is to predict the label of the target activity in t_a time steps ahead of when it occurs, by observing a video clip with a length of t_o time-steps. Thus, the anticipation task can be formulated as follows. Given an observed video clip $V_o = \{v_1, v_2, ..., v_{t_o}\}$ containing t_o segments, after anticipating the next content involving t_a time-steps, the model is required to predict the category of the activity at the current time-step. Here, each time-step spans a unit segment of τ seconds. As shown in Fig. 2, the proposed method consists of three key parts, i.e., observation summary, recurrent representation forecasting with memory bank, and activity anticipation with one-shot transferring, which will be introduced in detail in the following subsections.

3.2. Observation Summary

For egocentric activity anticipation, summarizing useful visual information from the observed video clip is crucial for the subsequent representation forecasting. Specifically, for each time-step i $(i = 1, 2, ..., t_o)$, we employ a shared feature extractor $\phi(\cdot)$, e.g., the I3D [5] or TSN [48] networks, to obtain the segment embeddings $\mathbf{e}_i \in \mathbb{R}^d$, as follows:

$$\mathbf{e}_i = \phi\left(v_i\right). \tag{1}$$

Then, we further leverage a recursive aggregator $\psi(\cdot)$, e.g., the GRU model, to sequentially integrate the segment embeddings $\mathbf{e}_i$ into a context feature $\mathbf{f}_{t_o} \in \mathbb{R}^d$ at the last observed time-step, as follows:

$$\mathbf{f}_{t_o} = \psi\left(\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{t_o}\right). \tag{2}$$

3.3. Recurrent Representation Forecasting with Memory Bank

Memory-augmented Representation Forecasting. In order to maintain prototypical activity semantics and explore the multi-hypotheses nature of the anticipation task, we introduce an external memory bank into the process of recurrent representation forecasting. Concretely, the memory bank is represented as a matrix $\mathbf{M} \in \mathbb{R}^{K\times d}$, where each

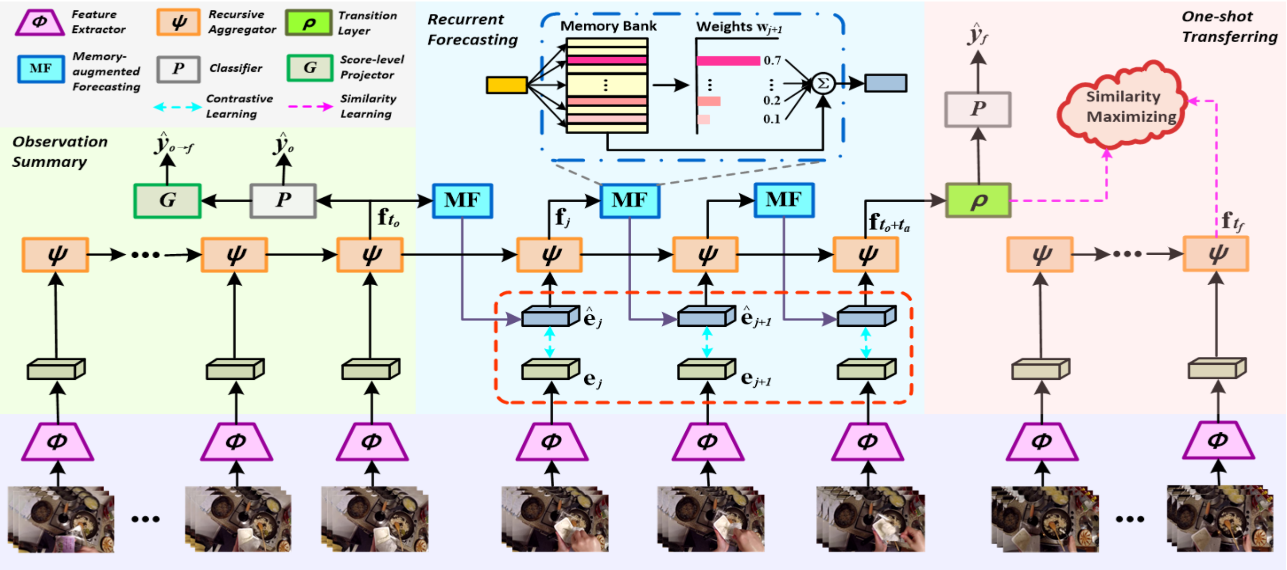


Figure 2. Overview of the proposed HRO method, which integrates memory-augmented recurrent and one-shot representation forecasting into a unified framework.

row is a memory item $\mathbf{M}_k$ $(k = 1, 2, , K)$ with feature dimension of d, and K is the memory size. In the recurrent representation forecasting stage, given the context feature $\mathbf{f}_j \in \mathbb{R}^d$ (i.e., a query) at time-step j $(j \geq t_o)$, the representation at the next time-step is predicted by an attention-based memory addressing mechanism, as follows:

$$\hat{\mathbf{e}}_{j+1} = \mathbf{w}_{j+1}\mathbf{M} = \sum_{k=1}^{K} w_{j+1}^{k}\mathbf{m}_k, \tag{3}$$

$$w_{j+1}^{k} = \frac{\exp\left(d\left(\mathbf{f}_j, \mathbf{m}_k\right)\right)}{\sum_{l=1}^{K}\exp\left(d\left(\mathbf{f}_j, \mathbf{m}_l\right)\right)}, \tag{4}$$

where $\mathbf{w}_{j+1} \in \mathbb{R}^{1\times K}$ is an attention weight vector normalized by a softmax operation, and the k-th entry of $\mathbf{w}_{j+1}$, i.e., w_{j+1}^{k}, indicates the contribution of the k-th memory item for forecasting representation at time-step $j+1$. $d\left(\cdot,\cdot\right)$ is a cosine similarity function, defined as follows:

$$d\left(\mathbf{f}_j, \mathbf{m}_k\right) = \frac{\mathbf{m}_k\mathbf{f}_j}{\left\|\mathbf{m}_k\right\|\left\|\mathbf{f}_j\right\|}. \tag{5}$$

Discussion on Memory Bank. Since the memory bank is sharable for the whole data sets during the training stage, it can summarize and store long-term prototypical activity semantics. Thus, at each time-step, the query context feature $\mathbf{f}_j$ can recall compact prototypical activity patterns from the memory bank to forecast discriminative future representation. In addition, from the multi-hypotheses modeling perspective, each memory item $\mathbf{m}_k$ can be regarded as a kind of potential hypothesis. Considering the uncertain nature of future anticipation, a distribution, i.e., $\mathbf{w}_{j+1}$, is derived to reflect the probability of each hypothesis being future.

Contrastive Learning. To further regularize the learning process of recurrent representation forecasting, motivated by the works in [40, 54], we employ a contrastive learning paradigm, which improves the representational ability by distinguishing the positive pairs among those negatives. Specially, at each forecasting time-step j, given the target positive sample $\mathbf{e}_j$ and negative sample set $S = \left\{\mathbf{s}_j^n\right\}_{n=1}^{N}$ with N samples, the contrastive loss function is defined as follows:

$$\mathcal{L}_j^{con} = -\log\frac{\exp\left(\varphi\left(\hat{\mathbf{e}}_j, \mathbf{e}_j\right)\right)}{\exp\left(\varphi\left(\hat{\mathbf{e}}_j, \mathbf{e}_j\right)\right) + \sum_{\mathbf{s}_j^n\in S}\exp\left(\varphi\left(\hat{\mathbf{e}}_j, \mathbf{s}_j^n\right)\right)}, \tag{6}$$

where $\varphi\left(\cdot\right)$ is a dot product function and $\hat{\mathbf{e}}_j$ is the anticipated representation. The target positive sample $\mathbf{e}_j$ is generated by extracting features using the shared feature extractor $\phi(\cdot)$ at the current forecasting time-step, i.e., $\mathbf{e}_j = \phi\left(v_j\right)$. To guarantee the diversity of the negative sample set, we randomly sample features from the video clips, which have different activity labels from the target positive sample. Minimizing this contrastive loss forces the model to be aware of semantic differences between different egocentric activities, which is beneficial to anticipating discriminative future representations.

Memory Bank Learning. Empirically, we find that it is insufficient to learn a compact yet discriminative memory bank, by merely relying on the classification loss for the activity anticipation task. Therefore, we devise two new loss terms, based on the diversity and sparsity schemes, to regulate the learning of the memory bank. First, the diversity scheme implies that each memory item corresponds to a unique activity prototype, which should be distinguished from other items. Thus, the diversity loss is formulated based on the orthogonality constraint, as follows:

$$\mathcal{L}^d = \left\| \mathbf{M}\mathbf{M}^{\mathrm{T}} - \mathbf{I} \right\|_F , \tag{7}$$

where $\mathbf{I} \in \mathbb{R}^{K \times K}$ is an identity matrix, and $\|\cdot\|_F$ denotes the Frobenius norm. Second, it is reasonable to predict the future representation at a specific time-step, by recalling a small portion of items from the memory bank, rather than densely sampling prototypical activity patterns from all of the memory items. Therefore, to avoid assigning a uniform probability distribution over the K hypotheses (i.e., memory items), we formulate a sparsity regularization loss to impose a constraint on the weight vector $\mathbf{w}_j$, at each forecasting time-step, as follows:

$$\mathcal{L}_j^s = \left\| \mathbf{w}_j \right\|_1 . \tag{8}$$

3.4. Activity Anticipation with One-shot Transferring

One-shot Transferring. Since the egocentric activity anticipation is essentially an online task, most of the current methods mainly rely on the observed information and forecasted representations before the target anticipation time-step. However, even though the forecasted representations contain correct activity semantics, we observe that it is still difficult to achieve satisfactory anticipation performance, due to the lack of contextual features after the anticipation time-step. Thus, to mitigate this issue, we propose to explicitly explore the relations between the representations before and after the anticipation time-step in an offline manner, during the training phase. To avoid long period of recurrent forecasting, we leverage a one-shot transferring paradigm, by introducing a representation transition layer, to bridge the gap between the semantics before and after the target anticipation time-step.

Specifically, in the offline training, given the video clip after the anticipation time-step denoted as $V_f = \left\{ v_{t_o+t_a+1}, ..., v_{t_f} \right\}$, we first summarize its content by extracting context feature $\mathbf{f}_{t_f}$, i.e., $\mathbf{f}_{t_f} = \psi\left(\phi\left(v_{t_o+t_a+1}\right), ..., \phi\left(v_{t_f}\right)\right)$. Then, we introduce a transition layer $\rho\left(\cdot\right)$, which projects the anticipated context feature $\mathbf{f}_{t_o+t_a}$ into another space to simulate the representations containing the future activity semantics of V_f, in a one-shot transferring manner. To achieve this goal, we define a feature similarity learning loss, as follows:

$$\mathcal{L}^T = \exp\left(-d\left(\rho\left(\mathbf{f}_{t_o+t_a}\right), \mathbf{f}_{t_f}\right)\right) . \tag{9}$$

where $d\left(\cdot,\cdot\right)$ is a cosine similarity measurement function, and the transition layer $\rho\left(\cdot\right)$ is implemented as a multi-layer perceptron (MLP).

Training Objective. The representation generated by the one-shot transfer is directly utilized for egocentric activity anticipation at the target time-step, i.e., $\hat{y}_f = P\left(\rho\left(\mathbf{f}_{t_o+t_a}\right)\right)$, where $P\left(\cdot\right)$ denotes a linear classifier and $\hat{y}_f$ is the anticipated activity category. To facilitate the learning of the basic feature extractor and aggregator, we also impose an activity classification task on the observed video clip, using the shared classifier, i.e., $\hat{y}_o = P\left(\mathbf{f}_{t_o}\right)$. In addition, considering that two consecutive activities in a long period of an event usually involve latent semantic relations, we mine this prior knowledge by explicitly exploring the score-level transition. Thus, we employ another linear projector $G\left(\cdot\right)$ to predict the next activity, based on the predicted activity score of the observed clip, i.e., $\hat{y}_{o\to f} = G\left(\hat{y}_o\right)$. Thus, the classification-level loss function is designed by integrating these three parts, as follows:

$$\mathcal{L}^C = \mathcal{L}_{CE}\left(\hat{y}_f, y_f\right) + \mathcal{L}_{CE}\left(\hat{y}_o, y_o\right) + \mathcal{L}_{CE}\left(\hat{y}_{o\to f}, y_f\right) , \tag{10}$$

where $\mathcal{L}_{CE}\left(\cdot,\cdot\right)$ denotes the cross-entropy loss function, y_f and y_o are the ground-truth label of target future activity and observed activity, respectively. By incorporating other learning criteria in Eq. (6), Eq. (7), Eq. (8) and Eq. (9), the overall training objective function is defined as follows:

$$\mathcal{L} = \mathcal{L}^C + \alpha \sum_{j=1}^{t_a} \mathcal{L}_j^{con} + \beta \mathcal{L}^d + \gamma \sum_{j=1}^{t_a} \mathcal{L}_j^s + \lambda \mathcal{L}^T , \tag{11}$$

where the parameters α, β, γ and λ are used to balance the contribution of the corresponding loss terms.

Inference. In the inference stage, we emphasize that our proposed framework only has access to the observed video clip (i.e., with a length of t_o time-steps), which follows the online manner for the egocentric activity anticipation task. Specifically, after extracting context features from the observed video clip, we first recursively forecast the representations using the memory bank, for t_a time-steps. Then, at the target anticipation time-step, the anticipated features are fed into the transition layer to obtain the future-semantics-enhanced representations, which are further utilized for predicting the target activity category score $\hat{y}_f$. On the other hand, we can predict the observed activity category score $\hat{y}_o$, based on the observed video clip. Then, we directly predict the category score of the next activity, $\hat{y}_{o\to f}$, using the score-level transition. Thus, the final category score of the target activity is obtained by averaging the scores of both $\hat{y}_f$ and $\hat{y}_{o\to f}$, i.e., $\hat{y}_{final} = \frac{\hat{y}_f + \hat{y}_{o\to f}}{2}$.

4. Experiments

4.1. Experimental Setup

Data Sets. We evaluate our proposed method on two large-scale egocentric activity data sets, including EPIC-Kitchens [9] and EGTEA Gaze+ [32]. The EPIC-Kitchens data set is constructed by collecting activities performed by 32 participants in diverse kitchen environments. It involves 125 verb

Table 1. Egocentric activity anticipation results of different methods on the EPIC-Kitchens validation set.

Methods	Top-5 Activity Accuracy (%) @ different t_a (s)								Top-5 Acc. (%) @1s			M. Top-5 Rec. (%) @ 1s		
	2	1.75	1.5	1.25	1	0.75	0.5	0.25	Verb	Noun	Act.	Verb	Noun	Act.
DMR [47]	/	/	/	/	16.86	/	/	/	73.66	29.99	16.86	24.50	20.89	03.23
ATSN [8]	/	/	/	/	16.29	/	/	/	77.30	39.93	16.29	33.08	32.77	07.60
MCE [15]	/	/	/	/	26.11	/	/	/	73.35	38.86	26.11	34.62	32.59	06.50
SVM-TOP3 [2]	/	/	/	/	25.42	/	/	/	72.70	38.41	25.42	41.90	34.69	05.32
ED [17]	21.53	22.22	23.20	24.78	25.75	26.69	27.66	29.74	75.46	42.96	25.75	41.77	42.59	10.97
FN [11]	23.47	24.07	24.68	25.66	26.27	26.87	27.88	28.96	74.84	40.87	26.27	35.30	37.77	06.64
RL [36]	25.95	26.49	27.15	28.48	29.61	30.81	31.86	32.84	76.79	44.53	29.61	40.80	40.87	10.64
EL [26]	24.68	25.68	26.41	27.35	28.56	30.27	31.50	33.55	75.66	43.72	28.56	38.70	40.32	08.62
RU-LSTM [16]	29.44	30.71	32.33	33.41	35.32	36.34	37.37	38.98	79.55	51.79	35.32	43.72	49.90	15.10
SRL [40]	30.15	31.28	32.36	34.05	35.52	36.77	38.60	40.49	/	/	35.52	/	/	/
LAI [54]	/	/	32.50	33.60	35.60	36.70	38.50	39.40	/	/	35.60	/	/	/
ActionBanks [41]	/	/	/	/	/	/	/	/	80.00	52.80	35.60	/	/	/
AVT [18]	/	/	/	/	/	/	/	/	79.90	54.00	**37.60**	/	/	/
HRO (Ours)	**31.30**	**32.67**	**34.26**	**35.87**	**37.42**	**38.36**	**39.89**	**42.36**	**81.53**	**54.51**	37.42	**45.16**	**51.78**	**17.50**

classes, 331 noun classes, and 2,513 unique activity classes, in total. We follow the same experimental setting in [16], where the 28,472 activity segments in the public training set are further split into 23,493 segments and 4,979 segments for training and validation, respectively. The EGTEA Gaze+ data set consists of 10,325 activity segments, involving 19 verb classes, 51 noun classes and 106 unique activity classes. We report the average performance over the three splits by following the evaluation setup in [32].

Implementation Details. For both the EPIC-Kitchens and EGTEA Gaze+ data sets, each time-step occupies a temporal field with 0.25 seconds, i.e., $\tau = 0.25$. Following the settings in [16, 40], the length of the basic observed time-steps is set to 6, i.e., $t_o = 6$, and the maximum length of the anticipated time-steps is set to 8, i.e., $t_a = 8$. Thus, the egocentric activity anticipation task is specified as follows. Based on the observation of a video clip with $1.5s$, the model is required to output the anticipation results at next eight time-steps, i.e., $0.25s$, $0.5s$, $0.75s$, $1s$, $1.25s$, $1.5s$, $1.75s$ and $2s$. For a fair comparison with other state-of-the-art methods, we directly utilize the features extracted from each time-step provided by [16]. The recursive aggregator $\psi(\cdot)$ is implemented with a GRU block, and the dimension of the hidden state is set to 1024, i.e., $d = 1024$. The memory bank sizes K for EPIC-Kitchens and EGTEA Gaze+ are set to 500 and 100, respectively. During the training stage, the length of the sampled video clip after the target anticipation time-step is $3s$, i.e., 12 time-steps. In the contrastive learning, at each time-step, we sample 128 negative samples to form the negative training set, i.e., $N = 128$. The weight parameters α, β, γ and λ in Eq. (11) are set as 0.7, 0.01, 0.02 and 0.01, respectively. Our proposed model is trained using the SGD optimization algorithm for 150 epochs, with an initial learning rate of 0.1 and a momentum of 0.9. The batch size is 128. For the multimodal evaluation setting, we first train three independent branches by individually taking RGB, optical-flow and object features as inputs. Then, the final anticipation results are obtained by a late fusion of the predictions from these three branches. For fair comparison, in the subsequent experiments, the results under the multimodal setting are reported, if not specified.

Table 2. Egocentric activity anticipation results of different methods on the EGTEA Gaze+ data set.

Methods	Top-5 Activity Accuracy (%) @ different t_a (s)							
	2	1.75	1.5	1.25	1	0.75	0.5	0.25
DMR [47]	/	/	/	/	55.70	/	/	/
ATSN [8]	/	/	/	/	40.53	/	/	/
MCE [15]	/	/	/	/	56.29	/	/	/
ED [17]	45.03	46.22	46.86	48.36	50.22	51.86	49.99	49.17
FN [11]	54.06	54.94	56.75	58.34	60.12	62.03	63.96	66.45
RL [36]	55.18	56.31	58.22	60.35	62.56	64.65	67.35	70.42
EL [26]	55.62	57.56	59.77	61.58	64.62	66.89	69.60	72.38
RU-LSTM [16]	56.82	59.13	61.42	63.53	66.40	68.41	71.84	74.28
LAI [54]	/	/	/	/	66.71	68.54	72.32	74.59
SRL [40]	59.69	61.79	64.93	66.45	70.67	73.49	78.02	82.61
HRO (Ours)	**60.12**	**62.32**	**65.53**	**67.18**	**71.46**	**74.05**	**79.24**	**83.92**

4.2. Comparison with State-of-the-art Methods

Results on EPIC-Kitchens. Table 1 tabulates the comparison results of our proposed method with other state-of-the-art approaches on the EPIC-Kitchens validation set. We can find that the proposed method consistently outperforms other competitors at all evaluated anticipation times. Our proposed HRO framework achieves the Top-5 activity accuracy of 37.42%, which outperforms ActionBanks [41] by 1.82% and is comparable to the performance of Transformer-based model AVT [18] (37.6%). Both ATSN [8] and MCE [15] directly generalize the classical activity recognition framework, i.e., TSN [48], for the anticipation task, which is insufficient to achieve satisfactory performance. The methods, which anticipate future activity without forecasting the unobserved content, e.g., FN [11], RL [36] and EL [26], perform worse than the recursive-anticipation-based frameworks, e.g., RU-LSTM [16], SRL [40] and LAI [54]. This suggests that compensating the missing semantics of the unobserved video parts is necessary. RU-LSTM employs rolling and unrolling LSTM blocks, which account for summarizing observations and predicting future activities, respectively. However, the vanilla recurrent-unit-based anticipation strategy is prone to accumulated errors in forecasting intermediate representation. To mitigate this issue, SRL and LAI introduce contrastive learning to regulate

Table 3. Ablation experimental results under different configurations with respect to memory bank learning on the EPIC-Kitchens validation set.

Exp.	$\mathcal{L}^{con}$	$\mathcal{L}^{d}$	$\mathcal{L}^{s}$	MB	Top-5 Activity Accuracy (%) @ different t_a (s)				
					2	1.25	1	0.75	0.25
1	✓	-	-	-	29.02	33.39	34.81	35.88	38.48
2	✓	-	-	✓	29.42	33.63	35.18	36.14	38.92
3	✓	-	✓	✓	29.65	33.77	35.36	36.36	39.34
4	✓	✓	-	✓	29.97	33.93	35.52	36.67	40.26
5	✓	✓	✓	✓	30.23	34.35	35.64	37.02	40.87

Table 4. Ablation experimental results under different configurations with respect to contextual semantics exploration on the EPIC-Kitchens validation set.

Config.	Top-5 Activity Accuracy (%) @ different t_a (s)							
	2	1.75	1.5	1.25	1	0.75	0.5	0.25
w/o FS	30.23	31.52	32.76	34.35	35.64	37.02	38.93	40.87
w/ FS (RF)	30.62	32.04	33.47	34.92	36.53	37.68	39.36	41.40
w/ FS (OT)	31.30	32.67	34.26	35.87	37.42	38.36	39.89	42.36

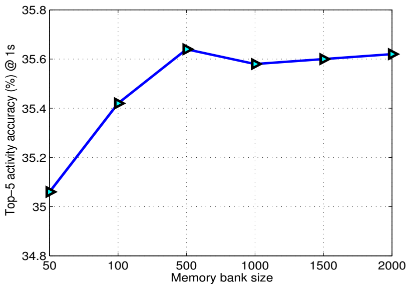


Figure 3. Comparison results in terms of Top-5 activity accuracy (%) at 1s under different memory bank sizes.

the process of future-representation forecasting in a self-supervised manner. Different from these existing recursive anticipation methods, our proposed framework learns an external memory bank, which helps to forecast discriminative representations, by maintaining long-term prototypical activity semantics. The averaged improvement over 1.5% on Top-5 activity accuracy for all anticipation times validates the effectiveness of the proposed hybrid recursive and one-shot representation forecasting framework. Moreover, in terms of the class-aware metric, i.e., Mean Top-5 Recall, our proposed HRO model outperforms the reported second-best results (RU-LSTM) by 1.44%, 1.88% and 2.4%, when anticipating the future verb classes, noun classes and activity classes, respectively.

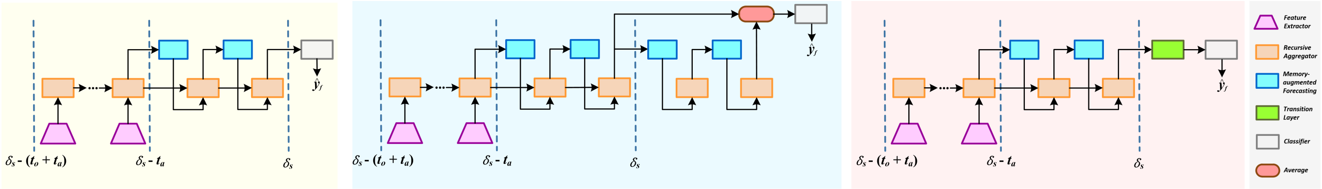
Results on EGTEA Gaze+. Table 2 presents the Top-5 activity accuracy results of different methods on the EGTEA Gaze+ data set, at eight evaluated anticipation time-steps. The proposed HRO method outperforms other state-of-the-art approaches at all evaluated anticipation times. Compared with the reported second best (SRL), our results are more than 0.77% higher, on average, with respect to the eight anticipation times. This, once again, reveals the robust performance of our proposed framework.

4.3. Ablation Study

Effect of the Memory Bank Learning. We conduct ablation experiments to explore the influence of the memory bank and two auxiliary loss terms (i.e., $\mathcal{L}^d$ and $\mathcal{L}^s$). The results are summarized in Table 3. For comparison, we implement a baseline model (Exp. 1), which employs vanilla GRU blocks to recursively forecast future representations, supervised by the contrastive learning. Based on the ablation results, we have the following observations: (1) Introducing the memory bank into recurrent representation forecasting can bring a consistent performance improvement at all activity anticipation times. Since the memory bank maintains long-term prototypical activity semantics, it can help the recurrent prediction model to reconstruct the most discriminative patterns at each anticipation time-step thus alleviating error accumulation issues, to some extent. (2) The sparsity regularization loss $\mathcal{L}^s$ is designed to force the anticipation model to concentrate on the most likely hypothesis from the memory bank. We can find that this results in an averaged performance gain of 0.24% on Top-5 activity accuracy, which demonstrate the necessity of the sparsity loss $\mathcal{L}^s$. (3) The diversity loss $\mathcal{L}^d$ aims to guarantee that each learned item in the memory bank is unique. Without the regularization of $\mathcal{L}^d$, some memory items may be redundant, which degrades the discriminative ability of the memory bank. It can be observed that an averaged performance improvement of 0.61% is obtained when using the diversity loss $\mathcal{L}^d$ to guide the memory bank learning. (4) These two loss terms, i.e., $\mathcal{L}^s$ and $\mathcal{L}^d$, make complementary contributions to the learning of the memory bank. By combining them together, the Top-5 activity accuracy is 35.64% at anticipation time of 1s, which is higher than applying them individually (35.36% and 35.52%).

Influence of the Memory Bank Size. The number of memory items is a crucial hyper-parameter, which may influence the representational capacity of the memory bank. Thus, we conduct ablation experiments using different memory bank sizes, and report the Top-5 activity accuracy at anticipation time of 1s in Fig. 3. Generally, when the memory bank size is small (e.g., $K = 50$), it is unable to score sufficient activity prototypes, which leads to an inferior performance. The performance can be improved by increasing the memory size. When the memory size continuously increases, the performance tends to be stable, as the memory bank has already learned sufficient prototypical activity semantics for future representation forecasting.

Effectiveness of the One-shot Transferring. As introduced in Sec. 3.4, in the offline training, we leverage contextual semantics to enhance the forecasted representations at the target anticipation time-step, via a one-shot transferring paradigm. To evaluate its contribution, as shown in Fig. 4, we conduct experiments by utilizing two baseline



(a) w/o FS

(b) w/ FS (RF)

(c) w/ FS (OT)

Figure 4. Structure comparison in terms of using different future-semantics exploration strategies. (a) Baseline model w/o FS. (b) Baseline model w/ FS (RF). (c) Our proposed method w/ FS (OT).

models, i.e., w/o FS (Fig. 4 (a)) and w/ FS (RF) (Fig. 4 (b)). The former baseline merely forecasts representations recursively until the target anticipation time-step, without exploring future semantics (FS). The latter baseline leverages the future semantics by recursively forecasting (RF) representations continuously after the anticipation time-step. The comparison results are tabulated in Table 4. We can find that exploiting contextual information, based on the single recursive anticipation model (w/ FS (RF)), can still bring an averaged improvement of 0.59% on Top-5 activity accuracy, which shows the benefits of exploring future semantics in the offline training. The proposed one-shot transferring strategy, i.e., w/ FS (OT) (Fig. 4 (c)), further improves the utilization efficiency of future semantics, which results in an averaged performance gain of 0.76% over all evaluated anticipation times. The underlying reason is that the error accumulation is the inherent limitation of the recurrent forecasting strategy when applied to long-period anticipation. In contrast, our proposed one-shot transfer can alleviate this issue by modeling the relations between past and future representations, via the holistic similarity learning, which enriches the forecasted features by injecting future activity semantics at the target anticipation time-step.

4.4. Qualitative Results

Figure 5 illustrates three examples of the activity anticipation results at four evaluated times, predicted by the single recursive (SR) baseline and our proposed HRO model. In the first (top) and third (bottom) examples, the SR baseline fails to detect the activity changes after 1.5s, as it is prone to accumulated errors in relatively long periods of anticipation. In contrast, the proposed HRO method can mitigate this issue with the help of memory-augmented contrastive learning and future-semantics-guided one-shot transferring. In the second example (middle), the SR baseline may be misled by the inactive object, i.e., "plate", in the field of view, thus resulting in anticipating the wrong activity class of "take plate". The moment at 1.5s in the third example is at the transition boundary of two consecutive egocentric activities, which makes our HRO model ambiguous by predicting either "cut sauce" or "put_down sauce". However, the proposed HRO method can still predict the correct activity class of "put_down sauce" at 2s, as the forecasted representations are recalibrated at each anticipation time-step via a transition layer.

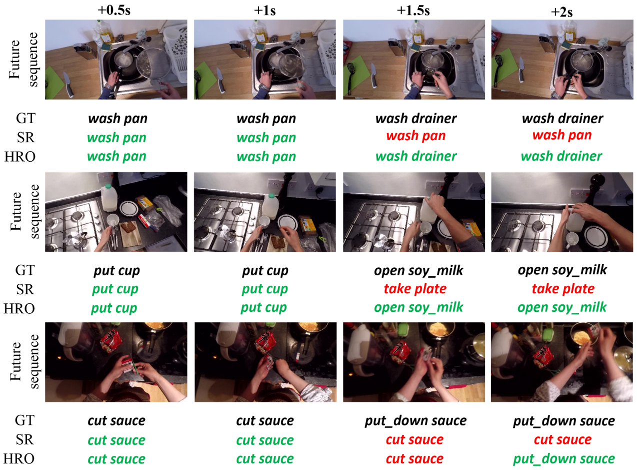


Figure 5. Visualization of the activity anticipation results, predicted by the single recursive (SR) baseline and our proposed HRO model, on the EPIC-Kitchens data set. The correct and incorrect results are indicated by green and red colors, respectively.

5. Conclusion

In this paper, we propose a hybrid egocentric activity anticipation framework by incorporating both recurrent and one-shot representation forecasting strategies, which are responsible for compensating the unseen activity semantics of the anticipation interval and after the anticipation moment, respectively. Specifically, a memory-augmented contrastive learning paradigm is presented to regularize the process of the recurrent representation forecasting, by penalizing the inaccurate anticipated features deviated from the prototypical activity semantics. Moreover, we propose a one-shot transferring paradigm to further recalibrate the forecasted representations at the anticipated moment, by distilling future semantics, via similarity learning in the offline training. Experimental results on two large-scale data sets demonstrate the superior performance of the proposed method.

References

[1] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 280–289, 2017. 3

[2] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Smooth loss functions for deep top-k classification. In *International Conference on Learning Representations (ICLR)*, 2018. 6

[3] Minjie Cai, Feng Lu, and Yoichi Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14380–14389, 2020. 1

[4] Yu Cao, Daniel Barrett, Andrei Barbu, Siddharth Narayanaswamy, Haonan Yu, Aaron Michaux, Yuewei Lin, Sven Dickinson, Jeffrey Mark Siskind, and Song Wang. Recognize human activities from partially observed videos. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2658–2665, 2013. 3

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020. 2

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 3

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 3, 6

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2021. 2, 5

[10] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *European Conference on Computer Vision (ECCV)*, pages 269–284, 2016. 3

[11] Roeland De Geest and Tinne Tuytelaars. Modeling temporal structure with lstm for online action detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1549–1557, 2018. 6

[12] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5343–5352, 2018. 1

[13] Alireza Fathi, Ali Farhadi, and James M. Rehg. Understanding egocentric activities. In *2011 International Conference on Computer Vision*, pages 407–414, 2011. 2

[14] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13224–13233, 2021. 3

[15] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018. 3, 6

[16] Antonino Furnari and Giovanni Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6251–6260, 2019. 1, 3, 6

[17] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. In *British Machine Vision Conference (BMVC)*, 2017. 6

[18] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13485–13495, 2021. 3, 6

[19] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton Van Den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1705–1714, 2019. 3

[20] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing Machines. *arXiv e-prints*, page arXiv:1410.5401, Oct. 2014. 3

[21] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 312–329, 2020. 3

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[24] Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, Jianhuang Lai, and Jianguo Zhang. Early action prediction by soft regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2568–2583, 2019. 3

[25] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 2

[26] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3118–3125, 2016. 6

[27] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[28] Georgios Kapidis, Ronald Poppe, and Remco C. Veltkamp. Multi-dataset, multitask learning of egocentric vision tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1

[29] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9917–9926, 2019. 3

[30] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3662–3670, 2017. 3

[31] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3054–3063, 2021. 3

[32] Yin Li, Miao Liu, and James M Rehg. 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)*, pages 619–635, 2018. 1, 2, 5, 6

[33] Yin Li, Zhefan Ye, and James M. Rehg. Delving into egocentric actions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287–295, 2015. 2

[34] Tianshan Liu, Kin-Man Lam, Rui Zhao, and Jun Kong. Enhanced attention tracking with multi-branch network for egocentric activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2021. 1

[35] Minghuang Ma, Haoqi Fan, and Kris M. Kitani. Going deeper into first-person activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903, 2016. 2

[36] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1942–1950, 2016. 6

[37] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 3

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[39] Guoliang Pang, Xionghui Wang, Jianfang Hu, Qing Zhang, and Wei-Shi Zheng. Dbdnet: Learning bi-directional dynamics for early action prediction. In *IJCAI*, pages 897–903, 2019. 3

[40] Zhaobo Qi, Shuhui Wang, Chi Su, Li Su, Qingming Huang, and Qi Tian. Self-regulated learning for egocentric video activity anticipation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1, 2, 3, 4, 6

[41] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision (ECCV)*, pages 154–171, 2020. 3, 6

[42] Yang Shen, Bingbing Ni, Zefan Li, and Ning Zhuang. Egocentric activity prediction via event modulated attention. In *Proceedings of the European conference on computer vision (ECCV)*, pages 197–212, 2018. 2

[43] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018. 2

[44] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9946–9955, 2019. 2

[45] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Learning to recognize actions on objects in egocentric video with attention dictionaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1

[46] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognitionn. In *British Machine Vision Conference (BMVC)*, 2018. 2

[47] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 98–106, 2016. 6

[48] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2019. 3, 6

[49] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3551–3560, 2019. 3

[50] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12249–12256, 2020. 2

[51] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 1

[52] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory Networks. *arXiv e-prints*, page arXiv:1410.3916, Oct. 2014. 3

[53] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 3

[54] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2021. 1, 2, 3, 4, 6