Learning to Anticipate Future with Dynamic Context Removal

Xinyu Xu[1], Yong-Lu Li[1,2], Cewu Lu[1] *
[1]Shanghai Jiao Tong University [2] Hong Kong University of Science and Technology
{xuxinyu2000, yonglu_li, lucewu}@sjtu.edu.cn

Abstract

Anticipating future events is an essential feature for intelligent systems and embodied AI. However, compared to the traditional recognition task, the uncertainty of future and reasoning ability requirement make the anticipation task very challenging and far beyond solved. In this filed, previous methods usually care more about the model architecture design or but few attention has been put on how to train an anticipation model with a proper learning policy. To this end, in this work, we propose a novel training scheme called ***Dynamic Context Removal*** *(**DCR**), which dynamically schedule the visibility of observed future in the learning procedure. It follows the human-like curriculum learning process,* i.e.*, gradually removing the event context to increase the anticipation difficulty till satisfying the final anticipation target. Our learning scheme is plug-and-play and easy to integrate any reasoning model including transformer and LSTM, with advantages in both effectiveness and efficiency. In extensive experiments, the proposed method achieves state-of-the-art on four widely-used benchmarks. Our code and models are publicly released at https://github.com/AllenXuuu/DCR.*

1. Introduction

Anticipating human action in the near future is a fundamental ability of humans as well as a basic requirement for intelligent systems with reasoning functionality. It serves as a support for many applications like autonomous driving [1, 40] and human-robot interaction [29, 42], where the future prediction of pedestrians and users are essential.

With the rapid evolution of deep learning techniques, the comprehensive understanding and analysis of human action videos attract attention in edging researches. In the traditional recognition field, modern video models [6, 13, 15, 34, 44, 48, 49, 53, 54] leverage spatiotemporal modeling to learn both spatial patterns and temporal logic and achieve significant progress in many video *recognition* tasks [8, 21, 26]. Besides, there is also growing interests in action *anticipation* [8, 9, 30, 33, 46]. Similarly, they both expect systems to discriminate the existing actions in videos. Differently, the observed video segment given for systems shifts forward in action anticipation, while in action recognition systems have the all the information of videos. Due to the temporal misalignment between visual observation and target action semantics, action anticipation is a much more challenging task than action recognition. It can hardly be simply treated as classification like video recognition for some reasons. First, the spatial configurations which deep neural networks (DNN) learned in the anticipation task is biased towards the supervision of future action labels, leading to the inaccurate representation of the current visual observation [18]. Second, the observation has a gap with the start time of action event, which challenges the high-level reasoning ability of model especially in the long-term dense action prediction setting [27, 43].

To tackle with the action anticipation, previous meth-

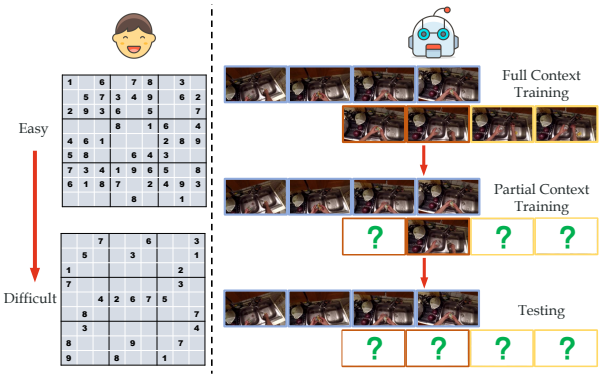

Figure 1. Revisiting learning curriculums in the classical Sudoku game, a kid starts with an easy Sudoku game of more observation (hints) then gets taught a harder level of less observable numbers. This reveals the curriculum learning process of how humans learn to reason in the physical world. In this work, we are inspired by learning Sudoku and build action anticipation model with similar curriculum designs. We leverage extra auxiliary frames in training but dynamically schedule their visibility to gradually strengthen the reasoning ability of model.

*Cewu Lu is the corresponding author, member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China and Shanghai Qi Zhi institute.

ods [10, 18–20, 22, 43, 57] proposed various neural architectures to focus on learning the temporal logic from past observations, with the intention to apply the past logic in reasoning the future. Though such methods achieve improvements, they still face performance bottlenecks on challenging benchmarks [8, 9, 30, 33, 46]. We argue that the reason is mainly that they did not learn from the way humans learn. In this work, we propose a simple but effective perspective for action anticipation. We want the model to learn temporal logic with the auxiliary of the *future snippet* but keep the functionality to reason out the future only given the observation in the past, which meets up the restriction of anticipation problem. To achieve our intention, we propose **D**ynamic **C**ontext **R**emoval **(DCR)** learning scheme, which integrates the motivation of curriculum learning [4] to train with sufficient context auxiliary at first then remove redundant context for better adaptation to a more difficult anticipation task, following the gradual learning process of humans. Fig. 1 gives an intuitive example.

Our training scheme is flexible and can easily advance different temporal reasoning architectures. Here, we mainly choose transformer [51] to implement our paradigm. First, in the *full context mode*, we propose the order-aware pre-training to learn video sequential order, which is a generalized method for transformer architecture. Next, in the *partial context mode*, we aim to reconstruct frames during action occurrence and dynamically schedule the visibility of auxiliary context. This learning paradigm conforms to how humans learn [4]. Apart from the transformer [51], we show our training scheme can also improve LSTM [24] based neural architecture.

We conduct experiments and analyses on four widely-used action anticipation benchmarks: EPIC-KITCHENS-100 [8], EPIC-KITCHENS-55 [9], EGTEA GAZE+ [33], 50-Salads [46]. Our training strategy turns out to be effective and achieves state-of-the-art on all four benchmarks. Moreover, we believe the proposed subtractive and adaptive paradigm can pave the way for the other complex and challenging temporal predictive tasks.

Our contribution includes: **(1)** We propose a novel learning scheme **DCR**, which advances the effectiveness and efficient of practical temporal modeling architecture including transformer [51] and LSTM [24]. **(2)** We propose a general order-aware pre-training for transformer architecture to carry out unsupervised pre-training using sequential order as supervision. **(3)** We achieve state-of-the-art on four widely-used action anticipation benchmarks.

2. Related Work

Action Anticipation is to predict action in the future by observing video clip with time τ_a before it occurs. It is required both in third-person [14, 19, 29, 30, 46, 52] and egocentric [8, 9, 17, 18, 20, 27, 33, 35, 43, 57] scenarios. It has a wide range of applications including intelligent robots [29, 42] and wearable devices. It used to have different task formulation such as dense action anticipation [43], but we consider to predict the next action [9, 18] in this work. Previous methods proposed various neural architectures including LSTM variants [14, 18, 19, 25, 57] and attention variants [20, 22, 43]. In the early work, Vondrick *et al*. [52] propose an unsupervised representation learning paradigm to connect the feature of present and future for the anticipation task. Li *et al*. [33] jointly model action anticipation with human gaze in egocentric videos. Later, Furnari *et al*. [18] propose a classic RULSTM architecture with modularity attention which achieves strong results. Sener *et al*. [43] attempt to anticipate action with different aggregations on the past. Some other works utilized extra knowledge like next active object [16] and hand motion [10] to anticipation action. One recent work AVT [20] leverages a causal transformer to model action anticipation in the *seq2seq* manner.

Video Sequential Order Modeling has been exploited in many tasks. Srivastava *et al*. [45] propose unsupervised learning techniques to learn generalized representation in video sequence. Zhou *et al*. [59] explore two simple tasks of pairwise ordering and future prediction in egocentric videos. Kong *et al*. [28] models sequential context relation to advance the recognition performance on part video observations. Misral *et al*. [39] propose a new perspective of video sequence as to verify whether the order is correct in learning. In our work, we leverage the permutation invariant property of self-attention and utilize sequential order as extra signals to perform self-supervised learning.

Vision Transformer gains much popularity recently, with a trend to exceed the classic convolution architecture in many visual tasks. Transformer [51] family originally raises in the language community, then permeates into the vision domain [12] including video related tasks [2, 13, 54]. It can be inserted as attention blocks [54, 56] into traditional video models as well as construct pure attention based video recognition architecture [2, 13]. In the filed of video action anticipation, transformer architecture can be directly used in temporal reasoning via causal attention [20].

Curriculum Learning is proposed by Bengio et al. [4]. It is motivated by the learning procedure of human, from easy to hard. It can be implemented via the schedule of category loss weights [31], data sampling [32] or other difficulty measurement [58]. This simple principle works well in many fields including language understanding [4], transfer learning [55] and more [31, 32, 58]. For the language reasoning task, previous work [7] also validates its effectiveness when doing baby-step short-term reasoning first. In our work, the richness of auxiliary context determines the easiness of the task. We schedule the context removal to obey the *easy-to-hard* principle of curriculum learning.

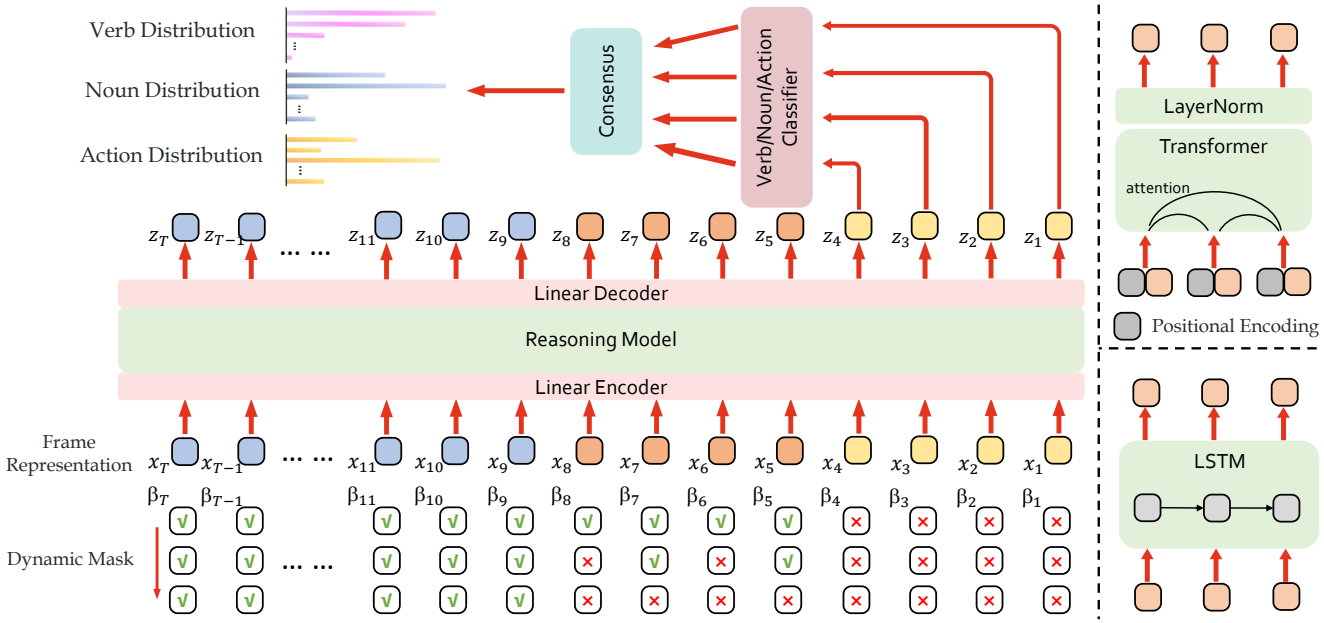

Figure 2. An overview of our training scheme. We intend to use past observations to reconstruct frames during action occurrence. The core of our motivation is to schedule the visibility in a curriculum learning manner, with more auxiliary frames at first but dynamically removed as the training goes on. The reconstructed action frames are sent to classifiers and make consensus to obtain final predictions. Our training scheme is flexible and can advance any reasoning models including attention-based transformer and traditional LSTM.

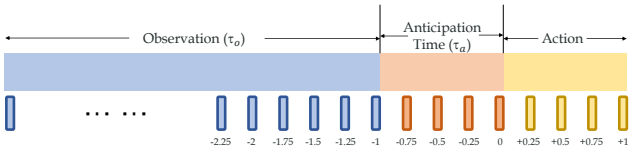

Figure 3. The general setting of the action anticipation task. Blue, red, yellow indicate different duration of past-observed, anticipation, action-occurring respectively.

3. Approach

We introduce the core of our work in this section. First, the detailed formulation of anticipation task is introduced in Sec. 3.1. Then, an overview of our method is described in Sec. 3.2. Our motivation is to adapt a well-trained model in the auxiliary context assistance setting into the anticipation setting by dynamic context removal. Thus, we leverage the order-aware pre-training (Sec. 3.3) to learn temporal dynamics for transformer [51] in *full context* mode. In Sec. 3.4, we describe the reconstruction driven curriculum design, which schedule the visibility of context and helps model tackle the anticipation problem gradually. Last, the learning objectives are described in Sec. 3.5.

3.1. Task Formulation

We briefly introduce our action anticipation setting in this section. As illustrated in Figure 3, there is a time gap between the observation and action segment. It is called as anticipation time, expressed as τ_a. We follow previous work [8,9,14,18,20,43,52,57] to fix τ_a = 1s on each benchmark. Another parameter is τ_o, which denotes the length of observed clip. Usually, τ_o is not restricted and any choice of τ_o is permitted. We sample frames at 4 fps following [18]. We use extra 8 frames ahead to assist training in our framework, but they are strictly not used in the validation and test.

3.2. Overview

We present an overview of our learning scheme in Figure 2. Assume we sample K frames for our model, then we start from K pre-extracted frame representations as $x_1, x_2, \cdots, x_K$, in the reverse chronological order. Each frame x_i is assigned with a binary mask $\beta_i \in \{0, 1\}$, determining its visibility. The mask is dynamically scheduled in different phase of training (introduced in Sec. 3.4), but we strictly set $\beta_1, \beta_2, \cdots, \beta_8 = 0$ in the test time. We project frame feature into a latent space, where a reasoning model $\mathcal{R}$ performs to reason out the masked frames based on visible information. Then, a linear decoder maps frames back to the original dimension. The goal of our reasoning model is to reconstruct the masked frames and we use $z_1, z_2, \cdots, z_K$ to denote the reconstruction. It is formulated as $z_1, z_2, \cdots, z_K = \mathcal{R}(x_1, \beta_1, x_2, \beta_2, \cdots, x_K, \beta_K)$. The last 4 frames $z_i (1 \leq i \leq 4)$ are frames in the action occurrence and they will be sent to the classifier to give prediction. For EPIC-KITCHENS series [8, 9] which also require marginalized verb/noun class prediction on their test server, we use verb/noun/action three classifiers on the top, but only apply single action classifier for other datasets. In the test time, predictions on these four frames are averaged

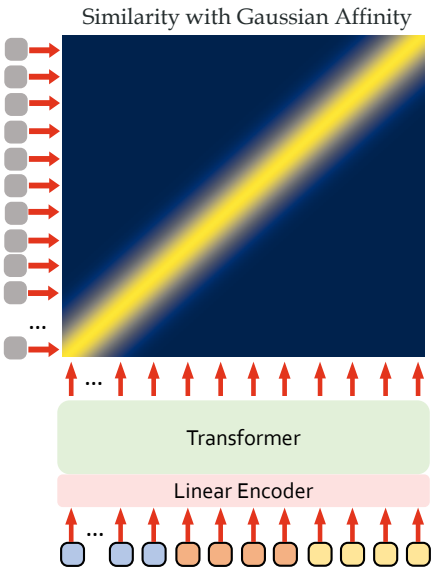

Figure 4. The order-aware pre-training for the permutation-invariant attention. We remove the positional encoding on the input side, but force the model to automatically understand the **order** of video sequence. It is trained by connecting the frame with its corresponding position to meet the pre-defined similarity.

to make a consensus [53] as the final result.

Noticeably, our training scheme is flexible and can be used for any reasoning models, including transformer [51], LSTM [24] *etc*. In this paper, we use transformer [51] for most experiments by default, but also give some LSTM [24] based results. A small difference of transformer and LSTM is about tackling masked frames. Mask is more practical for transformer based applications [11], so we direct assign zero value for the input. But for the recurrent LSTM structure, it's more sensitive about latest observation and zero value leads to the smooth prediction of future, thus we copy the masked frame using the latest visible (not masked) one.

3.3. Order-Aware Pre-training

For the transformer based reasoning model, we propose a novel order-aware pre-training to learn the temporal dynamics in the *full context mode*. In this stage, we notice that transformer is a permutation-invariant architecture without explicit positional encoding [51]. Thus, we use temporal position as signal to supervise the training and expect model to automatically recognize the order of input sequence, which implies the understanding of temporal logics among context. Such ordering based temporal modeling turns out to be effective in previous works [59].

We propose our self-supervised pre-training technique called order-aware pre-training. All frames, both the past observation and expanded 8 frames, are used in training. In another word, $\beta_i (1 \leq i \leq K)$ consistently equals 1. Without explicit integration of positional encoding, they are directly sent into the transformer after the linear encoder. Then we compute the cosine similarity between transformer output tokens and the positional encoding, which is followed by Softmax to probability space.

A pre-defined similarity label is required to supervise the training. The most naive choice is to use a diagonal matrix as similarity to treat it as a separate classification problem. However, time series is continuous, it would be much better to assign soft labels. To this end, we follow [23] to define similarity with Gaussian affinity. The similarity $s_{i,j}$ of positional encoding at time i and the frame feature at time j is measured as

$$s_{i,j} = \exp\left(-\frac{(i-j)^2}{\sigma^2}\right), \tag{1}$$

where σ is the bandwidth of Gaussian and we set $\sigma = 5$ in our experiments. Then the similarity is used to supervise the pre-training. We minimize the cross entropy loss with Gaussian affinity similarity as soft label.

The order-aware pre-training not only learns the relations in the context, but also provides a refinement of positional encoding. It's an aggregated context-agnostic representation on the whole dataset and more suitable for the context-removed scenario in the following training phase. Generally, the technique is motivated by the permutation-invariant property of self-attention and use sequence permutation as self supervised signals. We believe it can advance a wider range of transformer-based sequence modeling tasks, just like the success of masked language model [11].

3.4. Reconstruction Driven Curriculum Design

The goal of our expected anticipation model is to reconstruct masked future frames based on the visible context. In this stage, we use *partial context* in training and schedule the context visibility by the reconstruction quality. Here, we inherit the motivation of curriculum learning [4] as the system is given more auxiliary context at the beginning, referred as easy curriculum. We formulate the **easiness** of system as $T_e \in [0, 1]$, where e denotes the epoch in training. Specially, we have $T_1 = T_2 = 1$. Then, as the training goes on, we decrease the easiness of the task for our system and present the difficult anticipation task gradually.

In this phase, we consistently mask out the frames during action occurrence as $\beta_i = 0 (i \leq 4)$, which are colored yellow in Figure. 2. They are not directly for model input but serve as supervision of the reconstruction. This is also supported by our experiments that direct utilization of action frames harms classifier performance and reasoning model. On the contrary, past observation (blue frames in Figure. 2) are always visible at any time, as $\beta_i = 1 (i \geq 9)$. The median four frames (orange frames in Figure. 2) are the main field for designing different curriculums. They assist past observation to reconstruct anticipated action frames but

are dynamically removed determined by the easiness factor T_e. For $5 \leq i \leq 8$ We uniformly sample variable ρ_i in [0,1], as $\rho_i \in U(0,1)$. The frame x_i is visible only when ρ_i is smaller than the easiness factor T_e. It also means these frames have a probability T_e to be visible. It's formulated as $\beta_i = \mathbf{1}[T_e > \rho_i](5 \leq i \leq 8)$, where $\mathbf{1}[*]$ indicates the truth of statement and returns binary value. Generally, we obtain β series in Eq. 2:

$$\beta_i = \begin{cases} 1 & i \geq 9 \\ \mathbf{1}[T_e > \rho_i] & 5 \leq i \leq 8. \\ 0 & i \leq 4 \end{cases} \tag{2}$$

Empirically, we design a instance-specific local curriculum scheduling method in this work. Though a global schedule of T_e like linear or exponential may also work well in some scenarios (Sec. 4.6), we find it's sensitive on hyper-parameters tuning and not very convenient. To this end, we empirically apply an instance-specific easiness schedule. In each iteration, mask $\{\beta_1, \beta_2, \cdots, \beta_T\}$ is generated for a video clip. Assume $k(k \geq 5)$ is smallest for $\beta_k = 1$ then $x_1, \cdots, x_{k-1}$ are what we need to anticipate. We use the error of the 1-second future to measure the quality of reconstruction.

$$\begin{aligned} Q &= ||x_{k-4} - z_{k-4}||_2, \\ s.t. \quad k &= argmin[\beta_k = 1]. \end{aligned} \tag{3}$$

A **memory bank** is used to store the reconstruction quality for each case. It serves as criterion to define the *easiness* in next epoch. We have $T_1 = T_2 = 1$ at start, but simply schedule easiness T_e using the decline of Q in Eq. 4, with extra boundaries $\gamma_{min} = 0.95, \gamma_{max} = 1$ on the decreasing factor. In this case, rapid decline of Q represents a well learnt state of model for this case, thus we decrease easiness faster. The boundaries are used to stabilize easiness scheduling and guarantee the diversity of curriculums in different training stages.

$$\frac{T_e}{T_{e-1}} = \min\{\max\{\frac{Q_{e-1}}{Q_{e-2}}, \gamma_{min}\}, \gamma_{max}\}. \tag{4}$$

3.5. Learning Objective

We use two objectives to supervise the training process. One is about the predictive result of the next action class L_{cls}, while the other is the reconstruction loss of masked frames L_{rec}.

Prediction loss is used to supervise the prediction of 4 frames in the action segment. We adopt the cross entropy loss. In addition, we use label smoothing [47] techniques following [5] and find it works well in our task. This mainly attributes to the advantages of label smoothing on suppressing overfitting and maintaining uncertainty of future. Assume for Z_i, action classifier gives prediction $p_i^1, p_i^2, \cdots, p_i^C$, where C is the number of categories, Then, the action prediction loss L_{cls}^A can be formulated in Eq. 5, where y is the ground truth label, w_y is the class loss weight from class distribution and ϵ is the factor of label smoothing.

$$L_{cls}^A = \sum_{i=1}^{4} -(1-\epsilon) w_y \log(p_i^y) - \sum_{j=1}^{C} \frac{\epsilon}{C} \log(p_i^j). \tag{5}$$

For datasets which require marginalized verb/noun predictions additionally, we compute verb/noun prediction loss similarly as L_{cls}^V, L_{cls}^N. The prediction loss is made as $L_{cls} = L_{cls}^V + L_{cls}^N + L_{cls}^A$. For datasets only with an action classifier on the top, we have $L_{cls} = L_{cls}^A$.

Reconstruction loss is to teach our model reason out the masked frames based on the remaining context, just like the role of masked language prediction [11]. We expect the output representation of our reasoning model close to the original frame. Thus we simply use mean square error following [20] in Eq. 6 as feature-level supervision.

$$L_{rec} = \sum_{i=1}^{K} (1 - \beta_i) * ||z_i - x_i||_2. \tag{6}$$

Considering different scales and roles of two losses, we apply a weighted summation to obtain the total loss L_{total}:

$$L_{total} = \lambda_{cls} L_{cls} + \lambda_{rec} L_{rec}, \tag{7}$$

where λ_* are different weights for different loss items.

4. Experiment

In this section, we conduct comprehensive experiments and analysis on four widely used benchmarks to validate the effectiveness of our method.

4.1. Dataset and Metrics

EPIC-KITCHENS-100 (EK100) [8] is currently the largest dataset to support action anticipation task. It has 700 long videos of 100 hours about egocentric cooking activity. Each action class in EK100 consists of a verb and a noun. Totally, there are 97 verbs and 300 nouns, leading to 4,053 action compositions. There are 89,977 action segments whose labels are aggregated from unique narrations. The dataset splits into train/validation/test sets with the ratio of 75:10:15. The train and validation sets are publicly released but the test set is only able to be queried on the online server. The main metric for evaluation is recall@5, a class aware metric to avoid the long-tail bias of action distribution. Besides, the authors [8] also provide a tail action subset and an unseen participants subset to highlight the generalization performance of model.

EPIC-KITCHENS-55 (EK55) [9] is an earlier version of EK100. As a subset, it contains 432 videos in 55 hours.

	Method	Backbone	Verb	Noun	Action	# Params
RGB	RULSTM [18]	TSN	27.5	29.0	13.3	19.7M
	AVT [20]	TSN	27.2	30.7	13.6	303.9M
	AVT [20]	irCSN-152	25.5	28.1	12.8	409.6M
	AVT [20]	ViT*	28.7	32.3	14.9	383.8M
	DCR (LSTM)	TSN	27.9	28.0	14.5	14.1M
	DCR (LSTM)	TSM	28.4	28.5	15.2	20.2M
	DCR	TSN	31.0	31.1	14.6	78.2M
	DCR	TSM	**32.6**	**32.7**	**16.1**	84.3M
Flow	RULSTM [18]	TSN	19.1	16.7	7.2	19.7M
	AVT [20]	TSN	20.9	16.9	6.6	303.9M
	DCR (LSTM)	TSN	21.6	15.3	7.8	14.1M
	DCR	TSN	**25.9**	**17.6**	**8.4**	78.2M
Obj	RULSTM [18]	FRCNN	17.9	23.3	7.8	14.5M
	AVT [20]	FRCNN	18.0	24.3	8.7	298.8M
	DCR (LSTM)	FRCNN	16.1	19.6	7.5	10.1M
	DCR	FRCNN	**22.2**	**24.2**	**9.7**	74.2M

Table 1. Single branch results on EK100 [8] validation set. The backbone marked with * denotes end-to-end training.

	Method	Backbone	Top-1	Top-5	# Params
RGB	RULSTM [18]	TSN	13.1	30.8	18.5M
	ActionBanks [43]	TSN	12.7	28.6	112.9M
	AVT [20]	TSN	13.1	28.1	302.6M
	AVT [20]	ViT*	12.5	30.1	382.8M
	AVT [20]	irCSN-152	14.4	31.7	603.2M
	DCR	TSN	13.6	30.8	78.2M
	DCR	irCSN-152	15.1	**34.0**	82.0M
	DCR	TSM	**16.1**	33.1	82.0M
Flow	RULSTM [18]	TSN	8.7	21.4	18.5M
	ActionBanks [43]	TSN	8.4	19.8	112.9M
	DCR	TSN	**8.9**	**22.7**	78.2M
Obj	RULSTM [18]	FRCNN	10.0	29.8	13.2M
	ActionBanks [43]	FRCNN	10.2	29.1	52.5M
	DCR	FRCNN	**11.5**	**30.5**	74.2M

Table 2. Single branch results on EK55 [9] validation set [18]. The backbone marked with * denotes end-to-end training.

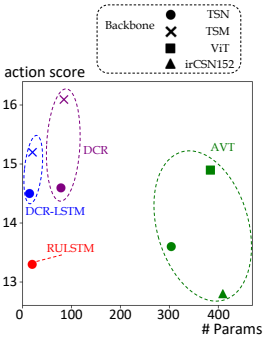

Figure 5. Score *vs*. Size.

There are 39,596 action segments, each assigned with a verb and noun class. It includes 125 verbs and 352 nouns in total. We follow the split of [9, 18]. The metric for evaluation is Top-1/5 accuracy.

EGTEA GAZE+ (EG+) [33] is another egocentric dataset for the joint modeling of action and gaze. We only use its action learning part. It contains 19 verbs, 51 nouns and 106 action compositions. There are 10,325 segments in 86 videos are annotated with action label. We report Top-5 accuracy and class-mean recall@5 over 3 standard official splits provided by the authors [33].

50-Salads (50S) [46] is a widely used third-person video dataset about salad preparation. It's a relatively smaller dataset then the previous ones as it only has nearly 0.9K action segments. And differently, its action class can't be marginalized into verb and noun. We follow [14, 20, 43] to use the 17-class coarse version of action annotation. label. We report Top-1 accuracy over 5 standard official splits provided by the authors [46].

Following previous works [8, 9, 14, 18, 20, 43, 52, 57], we set τ_a = 1s for all datasets. This setting is also shared for all baselines for a fair comparison.

4.2. Baseline

We compare DCR with respect to several competitive approaches including DMR [52], ATSN proposed in [9], MCE [17], FHOI [35], RULSTM [18], ActionBanks [43], ImagineRNN [57], Ego-OMG [10], AVT [20] and more. Please refer to supplementary material for more details about baselines.

4.3. Implementation Detail

Backbone. We adopt different types of feature (RGB appearance, Optical Flow and Object distribution) from different backbones. The first two modals can be encoded with (1) frame-level spatial model like ViT [12] or TSN [53] (2) snippet-level spatialtemporal model like TSM [34] or IG-65M pre-trained irCSN-152 [48]. Notably, for the spatialtemporal model $\mathcal{B}$ which requires k input frames, we use $x_t = \mathcal{B}\left(I_t, I_{t-1}, \cdots, I_{t-(k-1)}\right)$ to encode the feature, where I_t is the raw frame at time t. This avoids the involvement of future information. At last, feature of object distribution is represented by the category probability of all objects in the frame and we use the Faster-RCNN(FRCNN) [41] detector shared by [18]. Though prior art [20] uses a trainable backbone and benefits from the moving regularization of target frame, we consider more about efficiency and choose to freeze backbones in all experiments.

Observation. We set observation time τ_o = 10s for EPIC-KITCHENS [8, 9] series and 50S [46], but τ_o = 5s for EG+ [33]. Longer observation requirement is mainly because of the larger data scale for EPIC-KITCHENS [8, 9] and longer action duration in average for 50S [46].

Head Network. We use 6-layer, 16-head, 1024-dimensional transformer encoder model [51] optimized by AdamW [37] as the default reasoning architecture but also conduct experiments using 1-layer, 1024-dimensional LSTM [24] optimized by SGD [37] on EK100 [8]. We apply learning rate scheduling including 5-epoch warmup and half cosine annealing [36] in all experiments. Please refer to supplementary material for details about base learning rate, batch size, loss weight, *etc*.

4.4. Apples-to-Apples Comparison

We report uni-model performance as well as their trainable parameters on EPIC-KITCHENS series in Tab. 1,2 for a fair comparison. The baseline parameters are recorded from their public checkpoints. Our models have approximative parameters except different dimensionality of input space.

First, we report results on EK100 validation set in Tab. 1. On the most widely-used RGB-TSN backbone (in red), our LSTM version DCR is slightly lighter than classic RULSTM [18] while the transformer version is nearly quarter of AVT [20] (because of half network width). However, our

models consistently perform better, especially for the transformer version, which has 3.8%/1.0% performance gains over AVT on verb/action respectively. Additionally, a more effective TSM [34] backbone directly helps DCR to outperform the end-to-end trained AVT by 1.2% margin at lower expense. We scatter the performance and size of RGB-input models in Fig. 5. Apparently, our methods are in upper left corner, indicating advantages in both effectiveness and efficiency. Besides, on flow and obj modality, our DCR also outperforms previous works. Especially for the flow, we have 5.0% and 1.2% performance gains on verb and action respectively.

Next, for results on EK55 validation set in Tab. 2, our DCR with RGB-TSN backbone also exceeds all baselines (red) in a fair comparison. To our surprise, previous method [20] applies 12-layer deep transformer on irCSN-152 backbone to achieve best uni-model performance, but our light model easily outperforms it with 2.3% gain on Top-5 score (in blue). The stronger TSM backbone further improves top-1 action score by 1.7% over [20]. Besides, our method also achieves competitive results on flow and obj modalities.

Certainly, apples-to-apples comparisons verify the contribution of DCR in training effective anticipation model at lower expense. It clearly paves the way for further research.

4.5. Comparing to State-Of-The-Art

EPIC-KITCHENS. We late fuse different models to ensemble results on these two benchmarks. Despite previous work may use modality attention [18] or apply an extra transformer to aggregate multi-modal tokens [22], our last fusion results still show superiority in Tab. 3,4. On the validation set, we follow AVT [20] to use *rgb+obj* fusion and it outperforms baselines. For example, we have 1.7% performance gain on the whole EK100 and 3.6% top-5 action score on EK55. The competitions on the online leaderboard are more challenging. We make ensemble using models trained with *train+val* data. Our method outperforms previous works on most branches, except top-1 score on EK55 S2 test set of unseen participants. This is mainly because the competitive baseline Ego-OMG [10] adds delicate annotation of hand segmentation and active objects to learn intermediate knowledge representation, which helps anticipation in unseen environments. We argue results on leaderboard rely more on large-scale computation or extra training data. It doesn't matter to validate our effectiveness. For the details of our model ensemble and their weights, please refer to the supplementary material.

EGTEA GAZE+. We use TSN [53] feature on RGB and optical flow modalities following [18] to back this experiment. The final result is the late fusion of two branches in Tab. 5. Surprisingly, DCR has 1.5% and 2.5% performance gains over all baselines on top-5 accuracy and recall@5 respectively, establishing a new *state-of-the-art*.

Method	Validation			Test		
	Overall	Unseen	Tail	Overall	Unseen	Tail
RULTSM [18]	14.0	14.1	11.1	11.2	9.7	7.9
ActionBanks [43]	14.7	14.5	11.8	12.6	10.5	8.9
TransAction [22]	16.6	13.8	15.5	13.4	10.1	11.9
AVT [20]	15.9	11.9	14.1	16.7	12.9	13.8
DCR	**18.3**	**14.7**	**15.8**	**17.3**	**14.1**	**14.3**

Table 3. Result ensemble on EPIC-KITCHENS-100 [8].

Method	Validation		Test Seen (S1)		Test Unseen (S2)	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
ATSN [9]	-	16.3	6.0	28.2	2.3	9.4
ED [19]	-	25.8	8.1	18.2	2.4	6.6
MCE [17]	-	26.1	10.8	25.3	5.6	15.7
RULTSM [18]	15.3	35.3	14.4	33.7	8.2	21.1
FHOI [35]	10.4	25.5	15.4	34.3	8.6	22.9
ImagineRNN [57]	-	35.6	14.7	35.0	9.3	22.2
ActionBanks [43]	15.1	35.6	16.7	36.1	10.0	23.4
Ego-OMG [10]	**19.2**	-	16.0	34.5	**11.8**	23.8
AVT [20]	16.6	37.6	16.8	36.5	10.4	24.3
DCR	**19.2**	**41.2**	**17.7**	**38.5**	10.9	**24.8**

Table 4. Result ensemble on EPIC-KITCHENS-55 [9].

Method	Top-5	c.m. Recall@5
DMR [52]	55.7	38.1
ATSN [9]	40.5	31.6
MCE [17]	56.3	43.8
TCN [3]	58.5	47.1
ED [19]	60.2	54.6
RL [38]	62.7	52.2
EL [25]	63.8	55.1
RULSTM [18]	66.4	58.6
DCR	**67.9**	**61.1**

Table 5. Results on EG+ [33].

Method	Top-1
DMR [52]	6.2
RNN [14]	30.1
CNN [14]	29.8
ActionBanks [43]	40.7
AVT [20]	48.0
DCR	**51.1**

Table 6. Results on 50S [46].

50-Salads. Our training scheme is not limited to egocentric action anticipation, but also advances anticipation results in third-person videos. In this 3-rd view video benchmark, we use same ViT backbone to [20] and achieve 3.1% performance gain on top-1 accuracy score in Tab. 6.

4.6. Ablation Study

We conduct ablation study to verify the effects of our method on transformer [51] in Tab. 7, but leave LSTM [24]-based results in the supplementary material. We report on EK100 [8] and EG+ [33] with RGB inputs. (1) First, we compare a classification baseline by removing everything used in anticipation task. Each branch has a large performance drop, indicating basic classification technique is not suitable for direct anticipation. (2) Second, we consider model without order-aware pre-train. Their performance can't achieve best, especially 2.4% drop on EG+. (3) Third, we consider different easiness schedules. If we always train under T_e=1, it turns out that training and testing task have a large gap and model can't transfer well. If we always train without using future context as T_e=0, then the model gets trapped in local optimum and performs not very well. We consider different *global* schedules of T_e, like linearly decreases from 1 to 0 or exponentially multiplies $\gamma = 0.95$ after each epoch. These methods also bring advance in

	EK100 [8]		EG+ [9]
	TSM	TSN	TSN
DCR	**16.1**	**14.6**	**64.5**
classification	13.7	12.7	58.5
w.o. pre-training	15.5	14.3	62.1
$T_e = 1$	6.5	4.5	40.1
$T_e = 0$	15.2	13.8	62.9
linear T_e	15.0	13.9	64.0
exponential T_e	15.6	14.2	64.2
w.o. L_{rec}	13.5	12.6	56.0
w.o. label smooth	14.8	13.3	62.3

Table 7. Ablation study.

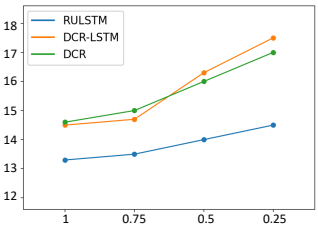

Figure 6. Effect of richer context via decreasing τ_a.

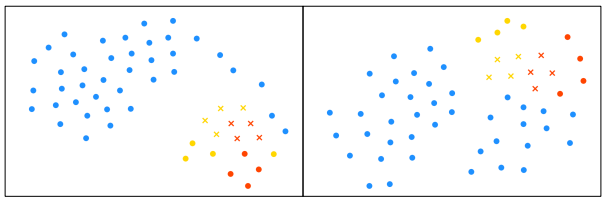

Figure 7. Qualitative cases of frame reconstruction. Blue dots are visible frames of past. Crosses are the predicted future representation, much closer to the actual future (yellow and red dots).

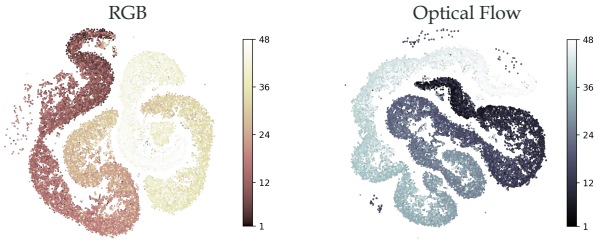

Figure 8. Effect of order-aware pre-training. We sample 1,000 segments from EK100 validation set and color the order-aware tokens from the pre-training models according to their temporal position. Our model can clearly embed the temporal dynamics within its learned manifold.

training the model, but are empirically worse than our *local* schedule proposition. (4) Last, we validate effects of loss components. Our model turns to have largest performance drop without L_{rec}, even worse than the classification. This is because different context complicates classification without feature-level supervision. Moreover, without label smoothing, we observe quick loss decreasing in training and worse performance due to overfitting.

4.7. Qualitative Results

We give qualitative results to better characterize the reasoning ability of our method.

First, we show what the model learns in the order-aware pre-training phase. We sample 1,000 segments from EK100 validation set, and extract output tokens from pre-trained order-aware transformer. In Fig. 8, we use t-SNE [50] to embed them into a 2D space and color frames according to their temporal position. It's an interesting visualization that indicating our pre-trained models have learnt video dynamics in the latent manifold, with a more comprehensive understanding of temporal logics in video.

Second, effect of auxiliary context shows in Fig. 6. We reduce anticipation time τ_a and test model performance. Thanks to our curriculum learning approach, our model retain stronger predictive ability in easy task. Compare to [18], our models have larger boost on more context. LSTM version DCR is more sensitive about latest context and even outperforms the transformer.

Last, we show qualitative cases of frame reconstruction. All frames and the model reconstruction are embedded via t-SNE [50] and visualized in Fig. 7. Blue ones are observed in anticipation task, while the yellow and red ones are the future. The reconstructed frames marked as cross are more close to the cluster of future frames.

5. Discussion

Limitations. We present an intuitive and empirical approach in video action anticipation, *e.g.* the local easiness scheduling. More finer-grained analysis in the future work is required to verify the curriculum learning approach.

Potential Negative Societal Impact. We train anticipation model on human-annotated datasets, which may import bias from human-defined labels. Due to the uncertainty of future and the potential bias, current anticipation model can hardly make robust prediction about the future action. But a possible solution for debiasing is to utilize unsupervised learning technique on a larger scale of data. On the usage of our method, video action anticipation technique is generally harmless except some malicious use for bad-intended events prediction. Thus, we encourage a proper use of technology that benefits mankind.

6. Conclusion

In this paper, we propose a novel strategy **DCR** on how to train an anticipation model. It follows the intuitive learning process of humans and flexibly advances any reasoning models in effectiveness and efficiency. In extensive experiments, we establish new *state-of-the-art* on four widely used benchmarks. We believe video action anticipation is an important problem for artificial intelligence, which supports many future applications. We are taking a great step in this field. In addition, our method is not limited to anticipation problem, but also has the potential to boost many other temporal predictive tasks. We hope our simple motivation of dynamic context removal can inspire more future works.

Acknowledgement

We appreciate the support from National Natural Science Foundation of China (No.72192821, 72192820), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and SHEITC (2018-RGZN-02046).

References

[1] Walter Morales Alvarez, Francisco Miguel Moreno, Oscar Sipele, Nikita Smirnov, and Cristina Olaverri-Monreal. Autonomous driving: Framework for pedestrian intention estimation in a real world scenario. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 39–44. IEEE, 2020.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[5] Guglielmo Camporese, Pasquale Coscia, Antonino Furnari, Giovanni Maria Farinella, and Lamberto Ballan. Knowledge distillation for action anticipation via label smoothing. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3312–3319. IEEE, 2021.

[6] 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 (CVPR)*, July 2017.

[7] Volkan Cirik, Eduard Hovy, and Louis-Philippe Morency. Visualizing and understanding curriculum learning for long short-term memory networks, 2016.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021.

[9] 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.

[10] Eadom Dessalene, Chinmaya Devaraj, Michael Maynord, Cornelia Fermuller, and Yiannis Aloimonos. Forecasting action through contact representations from first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2021.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers, 2021.

[14] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities, 2018.

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[16] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. Next-active-object prediction from egocentric videos. *Journal of Visual Communication and Image Representation*, 49:401–411, 2017.

[17] 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*, pages 0–0, 2018.

[18] Antonino Furnari and Giovanni Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[19] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017.

[20] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021.

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[22] Xiao Gu, Jianing Qiu, Yao Guo, Benny Lo, and Guang-Zhong Yang. Transaction: Icl-sjtu submission to epic-kitchens action anticipation challenge 2021. *arXiv preprint arXiv:2107.13259*, 2021.

[23] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

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

[25] 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. IEEE, 2016.

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola,

Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[27] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *CVPR*, June 2019.

[28] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2015.

[30] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 780–787, 2014.

[31] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23:1189–1197, 2010.

[32] Siyang Li, Xiangxin Zhu, Qin Huang, Hao Xu, and C-C Jay Kuo. Multiple instance curriculum learning for weakly supervised object detection. *arXiv preprint arXiv:1711.09191*, 2017.

[33] 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)*, September 2018.

[34] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[35] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *European Conference on Computer Vision*, pages 704–721. Springer, 2020.

[36] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[38] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1942–1950, 2016.

[39] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 527–544, Cham, 2016. Springer International Publishing.

[40] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Pedestrian action anticipation using contextual feature fusion in stacked rnns. In *BMVC*, 2019.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[42] M. S. Ryoo, Thomas J. Fuchs, Lu Xia, J. K. Aggarwal, and Larry Matthies. Robot-centric activity prediction from first-person videos: What will they do to me? In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 295–302, 2015.

[43] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding, 2020.

[44] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 1, 06 2014.

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

[46] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013.

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[48] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5551–5560, 2019.

[49] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018.

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[52] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016.

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[55] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning*, pages 5238–5246. PMLR, 2018.

[56] 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 (CVPR)*, June 2019.

[57] 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.

[58] Yunzhi Zhang, Pieter Abbeel, and Lerrel Pinto. Automatic curriculum learning through value disagreement. *Advances in Neural Information Processing Systems*, 33, 2020.

[59] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015.