Leveraging the Present to Anticipate the Future in Videos

Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, Du Tran

Leveraging the Present to Anticipate the Future in Videos

Antoine Miech[1,2] Ivan Laptev[1,2] Josef Sivic[1,2,3] Heng Wang[4] Lorenzo Torresani[4] Du Tran[4]
[1]Inria [2]École Normale Supérieure [3]CIIRC [4]Facebook AI

Abstract

Anticipating actions before they are executed is crucial for a wide range of practical applications including autonomous driving and robotics. While most prior work in this area requires partial observation of executed actions, in the paper we focus on anticipating actions seconds before they start. Our proposed approach is the fusion of a purely anticipatory model with a complementary model constrained to reason about the present. In particular, the latter predicts present action and scene attributes, and reasons about how they evolve over time. By doing so, we aim at modeling action anticipation at a more conceptual level than directly predicting future actions. Our model outperforms previously reported methods on the EPIC-KITCHENS and Breakfast datasets.

1. Introduction

Automatic video understanding has improved significantly over the last few years. Such advances have manifested in disparate video understanding tasks, including action recognition [5, 8, 11, 38, 41], temporal action localization [37, 39, 47, 53], video search [13], video summarization [32] and video categorization [29]. In this work, we focus on the problem of anticipating future actions in videos as illustrated in Figure 1.

A significant amount of prior work [5, 8, 11, 21, 23, 38, 41, 42, 46] in automatic video understanding has focused on the task of action recognition. The goal of action recognition is to recognize what action is being performed in a given video. While accurate recognition is crucial for a wide range of practical applications such as video categorization or automatic video filtering, certain settings do not allow for complete and even partial observation of action before it happens. For instance, an autonomous car should be able to recognize the intent of a pedestrian to cross the road much before the action is actually initiated in order to avoid an accident. In practical applications where we seek to act before an action gets executed, being able to anticipate the future given the present is critical.

Anticipating the future, especially long-term, is a challenging task because the future is not deterministic: several outcomes are possible given the current observation. To reduce uncertainty, most work in this field [2, 15, 18, 25, 35, 36] requires partially observed execution of actions. In this paper, we address the task of action anticipation even when no partial observation of the action is available. While prior work [7, 20, 27, 43] has addressed this same task, in this work, we specifically focus on better leveraging recognition models to improve future action prediction. We propose a fusion of two approaches: one directly anticipates the future while the other first recognizes the present and then anticipates the future, given the present. We have empirically observed complementary of these two approaches when evaluating on three distinct and diverse benchmarks: EPIC-KITCHENS [6], Breakfast [19] and ActivityNet [4].

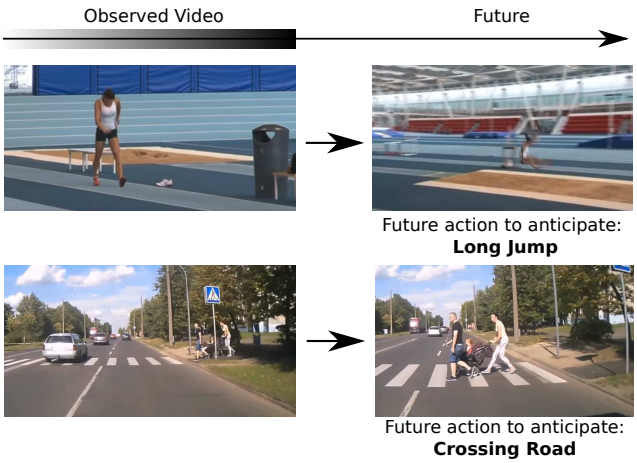

Figure 1: **Action anticipation**. Examples of action anticipation in which the goal is to anticipate future actions in videos seconds before they are performed.

1.1. Contributions

The contributions of our work are: **(i)** We propose a new framework for the task of anticipating human actions several seconds before they are performed. Our model is decomposed into two complementary models. The first, named the predictive model, anticipates action directly from the visual inputs. The second one, the transitional model, is first constrained to predict what is happening in the ob-

served time interval and then leverages this prediction to anticipate the future actions. **(ii)** We present extensive experiments on three datasets with state-of-the-art results on the EPIC-KITCHENS [6] and Breakfast action dataset [19]. In addition, our model provides ways to explain its ouputs, which allows us to easily interpret our model as we demonstrate in our qualitative analysis.

2. Related work

Predicting the future is a big area. Our work touches on future frame, motion and semantic mask prediction as well as human trajectory and action prediction, which we review below.

Future pixel, motion or semantic mask prediction. Future frame prediction has recently attracted many research efforts. Mathieu *et al*. [28] predict future frames of a video by proposing a multi-scale network architecture. They train it using an adversarial approach to minimize an image gradient difference loss. Vondrick *et al*. [44] also generate future video frames using a transformation of pixels from the past. Xue *et al*. [51] instead propose a probabilistic model to generate future frames from a single image. Oh *et al*. [30] predict action dependent future frames in old-school Atari video games. Finn *et al*. [10] explore video prediction for real-world interactive robot agents. Instead of directly predicting future pixels, Luc *et al*. [24] aim at predicting future semantic segmentation mask in videos, Walker *et al*. [45] explore future pose prediction in videos and Pinteal *et al*. [31] predict motion from single still images. Our work differs from them as we predict future action labels instead of predicting pixel level information.

Human trajectory prediction. Predicting human trajectories has also received wide attention [1, 26, 17, 50]. Kitani *et al*. [17] approach this task by casting it as an inverse reinforcement learning (IRL) problem. Alahi *et al*. [1] model prediction of human trajectories as a sequence generation task and propose to generate these trajectories using recurrent neural networks (LSTMs). These work differs from our as they are predicting an entire sequence of future locations instead of a single action.

Early-stage action recognition. Our work is related to the field of action anticipation and early-stage action prediction. A large body of work [2, 15, 18, 25, 35, 36] focuses on predicting actions given partially observed executions. This setting differs from action recognition [5, 8, 11, 21, 23, 38, 41, 42, 46] or temporal action detection [37, 39, 47, 53], as it assumes access to a small fraction (the beginning) of an action. One early work in this genre is from Ryoo [35], who uses dynamic bag-of-words to efficiently model the feature distribution change over time. Hoai *et al*. [15] and Ma *et al*. [25] formulate this task as a ranking problem where a monotonically non-decreasing prediction score is enforced as visual evidences are being accumulated. Similarly to Vondrick *et al*. [43], Shi *et al*. [36] aim first at regressing future visual feature vectors. These feature vectors are then used as input to an action recognition model for the early-stage action prediction. Our work differs from early-stage action prediction, as we aim at predicting an action even before it has actually started.

Action anticipation. Prior efforts [7, 17, 20, 27, 43] have been addressing the task of anticipating action before they are executed. The work of Farha *et al*. [7] aims at predicting not only one but a sequence of future actions. However, their experiments concern a restricted setup with a strong "action grammar" specific to cooking videos with predefined recipes [19]. Our work is not restricted to these type of datasets since we are also experimenting on unscripted cooking video dataset (EPIC-KITCHENS [6]) and non cooking video dataset (ActivityNet 200 [4]). Also their action anticipation approach can only be applied to videos with annotated sequence of actions whereas our method can be applied to any type of video dataset. Similarly to Mahmud *et al*. [27], their system is also trained to predict the start of the next action. Anticipating events before they occur is also used to predict traffic accidents [16, 40, 52]. Prior work also applied action anticipation in the domain of sports analytics such as basketball [3, 9], water polo [9], tennis and soccer [48]. Such models aim to anticipate future trajectories of a ball and individual players. IRL has also been recently applied to activity forecasting from first-person egocentric daily activity videos [33]. On the other hand, Wu *et al*. [49] combine on-wrist motion accelerometer and camera to perform daily intention anticipation. Note that several of these systems [1, 2, 7, 25, 27, 33, 36, 40, 49] employ the use of recurrent neural networks (RNNs) to address the sequential nature of these predictive tasks.

3. Action Anticipation Model

Our goal is to anticipate an action T seconds before it starts. More formally, let V denote a video. Then we indicate with $V_{a:b}$ the segment of V starting at time a and ending at time b, and with y_c the label of the action that starts at time c . We would like to find a function f such that $f(V_{0:t})$ predicts y_{t+T}. The main idea behind our model is that we decompose f as a weighted average of two functions, a predictive model f_{pred} and a transitional model f_{trans}:

$$f = \alpha f_{pred} + (1-\alpha) f_{trans},\ \alpha \in [0,1], \qquad (1)$$

where α is a dataset dependent hyper-parameter chosen by validation. The first function f_{pred} is trained to predict

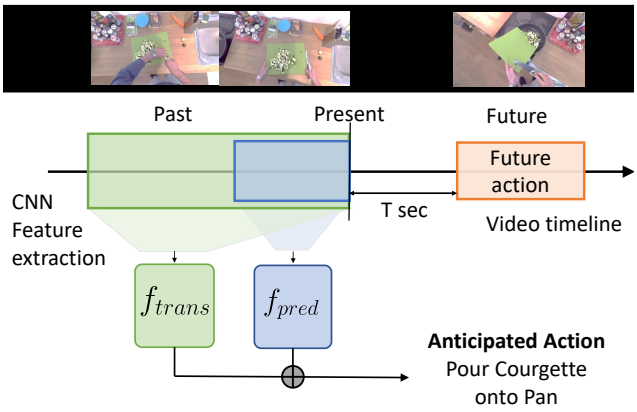

Figure 2: **Overview of our approach**. Our task is to predict an action T seconds before it starts to be performed. Our model is a combination of two complementary modules: the predictive model and the transitional model. While the predictive model directly anticipates the future action, the transitional model is first constrained to output what is currently happening. Then, it uses this information to anticipate future actions.

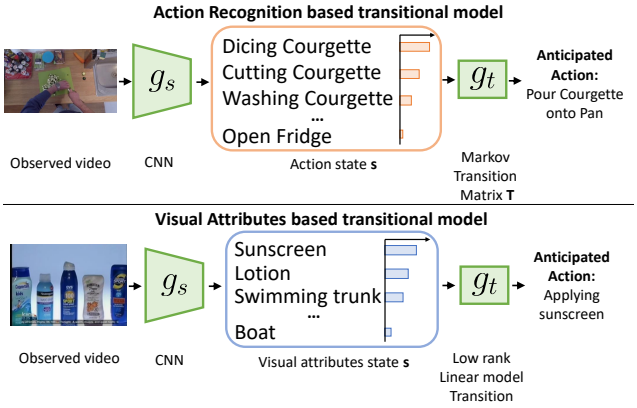

Figure 3: **Illustration of our transitional models**. Upper: our Action Recognition (AR) based transitional model learns to prediction future actions based on the predictions of an action recognition classifier applied on current/present frames (clips). Lower: our Visual Attributes (VA) based transitional model learns to predict future actions based on visual attributes of the current/present frames (clips).

the future action directly from the observed segment. On the other hand, f_{trans} is first *constrained* to compute high-level properties of the observed segment (e.g., attributes or the action performed in the present). Then, in a second stage, f_{trans} uses this information to anticipate the future action. In the next subsections we explain how to learn f_{pred} and f_{trans}. Figure 2 presents an overview of the proposed model.

3.1. Predictive model f_{pred}

The goal of the predictive model f_{pred} is to directly anticipate future action from the visual input. As opposed to f_{trans}, f_{pred} is not subject to any specific constraint. Suppose that we are provided with a training video V with action labels $y_{t_0+T}, \ldots, y_{t_n+T}$. For each label y_{t_i+T}, we want to minimize the loss:

$$l(f_{pred}(V_{s(t_i):t_i}),\ y_{t_i+T}), \tag{2}$$

where $s(t_i) = \max(0, t_i - t_{pred})$, l is the cross entropy loss, t_{pred} is a dataset dependent hyper-parameter, also chosen by validation, that represents the maximum temporal interval of a video f_{pred} has access to. This hyper-parameter is essential because looking too much in the past may add irrelevant information that degrades prediction performance. This loss is then summed up over all videos from the training dataset. In this work, f_{pred} is a linear model which takes as input a video descriptor which we describe in section 4.2.

3.2. Transitional model f_{trans}

The transitional model f_{trans} splits the prediction into two stages: g_s and g_t. The first stage g_s aims at recognizing a current state $\mathbf{s}$, describing the observed video segment. The state $\mathbf{s}$ can represent an action or a latent action-attribute. The second stage g_t takes as input the current state $\mathbf{s}$, and anticipates the next action given the current state $\mathbf{s}$. g_s can be thought of as a complex function extracting high-level information from the observed video segment, while g_t is a simple (in fact, linear) function operating on the state $\mathbf{s}$ and modeling the correlation between the present state and the future action. We will next explain in detail how we define the current state $\mathbf{s}$ and how we model the transition function g_t. We propose two different approaches for our transitional model: one that is based on action recognition and one that relies on visual attributes, as illustrated in Figure 3.

Transitional Model based on Visual Attributes. In this approach, we leverage visual attributes [23] to anticipate the future. Visual attributes have been previously used for action recognition by Liu *et al.* [23]. The idea is to first predefine a set of visual attributes describing the presence or absence of objects, scenes or atomic actions in a video. Then, a model is trained on these visual attributes for action recognition. In this work, we instead use visual attributes as a means to express the transitional model. The current state $\mathbf{s} \in [0,1]^a$ predicted by g_s, is then a vector of visual attributes probabilities, where a is the number of visual attributes. Given the presently observed visual attribute $\mathbf{s}$, g_t

predicts the future action. We model g_t as a low-rank linear model:

$$g_t(\mathbf{s}) = W_2(W_1\mathbf{s} + b_1) + b_2, \tag{3}$$

where $W_1 \in \mathbb{R}^{r\times a}, W_2 \in \mathbb{R}^{K\times r}, b_1 \in \mathbb{R}^r, b_2 \in \mathbb{R}^K$, $K \in \mathbb{N}$ the number of action classes and r is the rank of g_t. These parameters are learned, in the same manner as the predictive model, by minimizing the cross entropy loss between the predicted action given by $g_t(\mathbf{s})$ and the future action ground-truth. Implementing g_t through a low-rank model reduces the number of parameters to estimate. Empirically, we found that this leads to better accuracy, as shown in our experiments. The lower part of Figure 3 illustrates this case.

Transitional Model based on Action Recognition. Real-world videos often consist of a sequence of elementary actions performed by a person in order to reach a final goal such as *Preparing coffee*, *Changing car tire* or *Assembling a chair*. Many datasets come with a training set where each video has been annotated with action labels and segment boundaries for all occurring actions (e.g EPIC-KITCHENS, Breakfast). When this is available we can use action labels instead of predefined visual attributes for state $\mathbf{s}$. The intuition behind our claim is the fact that the anticipation of the next action significantly depends on the present being-performed action. In other words, we make a Markov assumption on the sequence of performed actions. More formally, suppose we are provided with an ordered sequence of action annotations $(a_0, \ldots, a_N) \in \{1, \ldots, K\}^N$ for a given video, where a_n defines the action class performed in video segment V_n. We propose to model $P(a_{n+1} = i|V_n)$ as follows:

$$P(a_{n+1} = i|V_n) = \sum_{j=1}^{K} P(a_{n+1} = i|\, a_n = j)P(a_n = j|V_n) \tag{4}$$

$\forall n \in \{0, \ldots, N-1\},\ i \in \{1, \ldots, K\}$. This reformulation decomposes the computation of $P(a_{n+1} = i|V_n)$ in terms of two factors: 1) an action recognition model $g_s(V_n)$ that predicts $P(a_n = j|V_n)$, i.e., the action being performed in the present; 2) a transition matrix T that captures the statistical correlation between the present and the future action, i.e., such that $T_{ij} \approx P(a_{n+1} = i|\, a_n = j)$. In this scenario, g_t takes as input the probability scores of each action given by g_s to anticipate the next action in a probabilistic manner:

$$g_t(\mathbf{s}) = T\mathbf{s}, \tag{5}$$

$$P(a_{n+1} = i) = \sum_{j=1}^{K} T_{i,j}s_j = [g_t(\mathbf{s})]_i\ . \tag{6}$$

In practice, we compute T by estimating the conditional probabilities between present and future actions from the the sequences of action annotations in the training set. The top part of Figure 3 illustrates this model.

Prediction Explainability. The transitional model f_{trans} provides interpretable predictions that can be easily analyzed for explanation. Indeed, the function g_t of the transitional model takes the form of a simple linear model applied to the state $\mathbf{s}$, both when using visual attributes as well as when using action predictions. The linear weights of g_t can be interpreted as conveying the importance of each element in $\mathbf{s}$ for the anticipation of the action class. For example, given an action class k to anticipate, we can analyze the linear weights of g_t to understand which visual attributes or action class are most responsible for the prediction of action class k.

It also provides an easy way to diagnose the source of mispredictions. For example, suppose the transitional model anticipates wrongly an action k and we seek to understand the reason behind such misprediction. Let $v_1, \ldots, v_a \in [0, 1]$ be the vectors encoding the visual attributes (or action recognition scores) for this wrong prediction. Let also $w_{k,1}, \ldots, w_{k,a} \in \mathbb{R}$ be the learned linear weights associated to the prediction of action class k. The top factor for the prediction of action k is $\max_{i\in[1,a]}(w_{k,i}v_i)$. By analyzing this top factor, we can understand whether the misprediction is due to a recognition problem (*i.e.* wrong detection score for the visual attribute/action class) or due to the learned transition weights.

4. Experiments

In this section, we evaluate our approach on three datasets. Then we provide an ablation study, compare our method with the state-of-the-arts and present qualitative analysis of the transitional model.

4.1. Datasets

These datasets were picked because they are diverse and contain accurate annotated action temporal segments necessary for the evaluation of action anticipation.

EPIC-KITCHENS. EPIC-KITCHENS [6] is a large-scale cooking video dataset containing 39,594 accurate temporal segment action annotations. Each video is composed of a sequence of temporal segment annotations. Three different tasks are proposed together with the dataset: object detection, action recognition and *action anticipation*. The action anticipation task is to predict an action one second before it has started. The dataset contains three different splits: the training set, the seen kitchens test set (S1) composed of videos from kitchens also appearing in the training

Model	Pretrain	Fine-tune	Action	Verb	Noun
ResNet-50	Imagenet	No	3.4	24.5	7.4
R(2+1)D-18	Kinetics	No	5.2	27.2	10.3
R(2+1)D-18	Kinetics	EK-Anticip.	5.0	24.6	9.7
R(2+1)D-18	Kinetics	EK-Recogn.	**6.0**	**27.6**	**11.6**

Table 1: **Effects of pre-training**. Action anticipation top-1 per clip accuracy on EPIC-KITCHENS with different models and pre-training datasets.

	Action		Verb		Noun	
	A@1	A@5	A@1	A@5	A@1	A@5
Transitional (VA)	4.6	12.1	25.0	71.7	9.1	24.5
Transitional (AR)	5.1	17.1	25.2	72.0	12.1	33.2
Predictive	6.3	17.3	27.4	73.1	11.9	31.5
Predictive + Transitional (VA)	**6.8**	18.1	**28.4**	**74.0**	12.5	33.0
Predictive + Transitional (AR)	6.7	**19.1**	27.3	73.5	**12.9**	**34.6**
Transitional (AR with GT)	16.1	29.4	29.3	63.3	30.7	44.4
Action recognition	12.1	30.0	39.3	80.0	23.1	49.3

Table 2: **Transitional and predictive model ablation**. Transitional model and predictive model ablation study on our EPIC-KITCHENS validation set with $T = 1$ *sec*. VA and AR denote for Visual Attributes and Action Recognition. Grey rows should be interpreted as accuracies upper bounds.

set and finally the unseen kitchens test set (S2) with kitchens that are not appearing in the training set. A publicly available challenge is also organized to keep track of the best performing approach on this anticipation task. Because of this public challenge, the labels of S1 and S2 test sets are not available. Thus, most of our results are reported on our validation set composed of the following kitchens: P03, P14, P23 and P30. We also report results evaluated by the challenge organizers on the held-out test set. Unless specified otherwise, for comparison purposes, we report experiments with $T = 1$ *sec*.

Breakfast. The Breakfast action dataset [19] is an annotated cooking video dataset of people preparing breakfast meals. It comes with 11267 temporal segment action annotations. Each video is also composed of a sequence of temporal action segment annotations. The dataset is partitioned into four different train / test splits: S1, S2, S3 and S4. We quantify performance with the average scores over all of the four splits. Unless specified differently, for comparison purposes, we report experiments with $T = 1$ *sec*.

ActivityNet 200. The ActivityNet 200 video dataset [4] contains 15410 temporal action segment annotations in the training set and 7654 annotations in the validation set. This video dataset is mainly used for evaluating action localization models but as the videos are provided accurate temporal segment for each action, we can also use them to evaluate models on action anticipation. As opposed to the EPIC-KITCHENS [6] and Breakfast [19] datasets, each video contains only one single action annotation instead of a sequence of action segments. For this reason, we cannot test on ActivityNet the transitional model based on action recognition. We only train and evaluate on videos in the datasets with at least 10 seconds of video before the action starts. In total, the training and validation sets consists of respectively 9985 and 4948 action localization annotations.

4.2. Video Representation

In this subsection we discuss how we represent the observed video segment V to perform action prediction. Our overall strategy is to split the video into clips, extract clips representation and perform pooling over these clips. Given an input video segment V, we uniformly split it into small clips $V = [V_1, \dots, V_N]$ where each clip V_i, $i \in [1, N]$ is short enough (*e.g.* 8 or 16 frames) that it can be fed into a pretrained video CNN C. From the penultimate layer of the CNN we extract an L_2-normalized one-dimensional representation $C(V_i)$ for each clip V_i. Then we perform a temporal aggregation $Agg([C(V_1), \dots, C(V_N)])$ of the extracted features in order to get a one-dimensional video representation for V. In our experiments, C is the R(2+1)D network of 18-layers from Tran *et al*. [41]. We perform a simple max pooling to aggregate features from all clips, but more sophisticated temporal aggregation techniques [29] can also be used in our model.

Visual Attributes. Our visual attributes presented in Section 3.2 include the taxonomies of Imagenet-1000 [34], Kinetics-600 [5] and Places-365 [54]. We train two ResNet-50 [12] CNN models: one on Imagenet-1000 and the other one on Places-365. For the Kinetics-600 taxonomy, we train a R(2+1)D-18 [41] model. In total, our set is composed of 1965 (1000+600+365) visual attributes. We densely extract these visual attributes every 0.5 seconds and apply the temporal max pooling operation to obtain a single vector for each video, as discussed above.

Leveraging the present for pretraining. In previous work [6, 43] the video representation was learned by finetuning a pretrained video CNN on the task of action anticipation. Instead, we propose to finetune the CNN representation on the task of action recognition on the target dataset. More specifically, instead of training the CNN on video clips sampled before action starts, we train it on clips sampled in the action segment interval. This is motivated by the fact that the task of action recognition is "easier' than action anticipation and thus it may lead to better feature learning. Table 1 reports accuracies on the EPIC-KITCHENS validation set obtained with our predictive model applied to

Model	Accuracy
Random baseline	0.3
Predictive	51.6
Transitional (All VA, Full rank)	48.0
Transitional (Object & Scene VA, Low rank, r = 256)	37.0
Transitional (All VA, Low rank, r = 256)	52.8
Predictive + Transitional (VA, Low rank)	**54.8**

Table 3: **Results on ActivityNet action anticipation**. Our methods compared with baseline on our validation set with $T = 5$ *sec.* VA stands for Visual Attributes.

different CNN representations. These results illustrate the benefit of fine-tuning the CNN on action recognition, instead of action anticipation as done in prior work [6, 43]. The Table provide also numbers for two additional baselines corresponding to 1) using the CNN pretrained on Kinetics without finetuning and 2) extracting features from a ResNet-50 2D CNN pretrained on Imagenet. It can be noted that the best accuracies for actions, verbs and nouns are obtained with the CNN finetuned on the action recognition task of EPIC-KITCHENS. Based on these results, in the rest of the work, we use CNN features computed from a R(2+1)D-18 first pretrained on Kinetics [5] and then finetuned for action recognition on the target dataset.

4.3. Ablation study

In order to understand the benefits of the different components in our model, we evaluate the predictive model separately from the transitional model. For the transitional model we report results for both the variant based on Visual Attributes (VA) as well as the version based on Action Recognition (AR). Table 2 summarizes the results achieved on the validation set of EPIC-KITCHENS [6]. The AR transitional model performs better than the VA transitional model. However, both are outperformed by the purely-predictive model. Interestingly, combining the predictive model with either of the two transitional models yields further accuracy gains. This suggests that the predictions are complementary.

We also show in grey, an accuracy upper bound achieved when directly recognizing the future frame as opposed to predicting from the past one (row Action recognition). The grey row Transitional (AR with GT) experiments shows the accuracy achieved when the transitional model is provided the groundtruth label of the last observed action. The improvement when using the groundtruth label is significant. This suggests that a large cause of missing performance is weak action recognition models and that better action recognition will produce stronger results for prediction.

We also perform ablation studies on the ActivityNet dataset in Table 3. Since we are not provided sequences of action annotations in this dataset, for this experiment we can only apply the transitional model based on Visual Attributes. Here again, we demonstrate the complementarity of the predictive and transitional models. The average of both approaches provides the best results for action anticipation. We also show the importance of modeling g_t as a low-rank linear model on visual attributes. Constraining g_t to be a low-rank linear model provides a boost of more than 4% in accuracy.

4.4. Comparison to the state-of-the-art

We compare our approach to the state-of-the-art on both the EPIC-KITCHENS and the Breakfast dataset. Table 5 shows our method compared to the recent work of Farha *et al*. [7]. The numbers for Vondrick *et al*. [43] are based on the reimplementation of this method provided in [7]. Table 4 reports results obtained from the EPIC-KITCHENS unseen kitchens action anticipation challenge submission server. Note that our EPIC-KITCHENS submission is done under the anonymous nickname of *masterchef* and is reported by the row **Ours (Predictive [D] + Transitional)** in this paper. On both datasets, our method outperforms all previously reported results under almost all metrics. Note that our best submitted model on the EPIC-KITCHENS challenge is simple and does not make use of any ensembling nor optical flow input.

4.5. Qualitative analysis

As explained in subsection 3.2, through the analysis of the transitional model f_{trans}, we can analyze which visual attributes are responsible for the anticipation of each action class. To do so, we analyze the linear weights from g_t (3) to list the top visual attributes maximizing the prediction of each action class. Table 6 shows some action classes from the ActivityNet 200 [4] dataset and the top-3 visual attributes that maximize their anticipation. For instance, we can observe that identifying a *Border collie* dog (A dog specialized in the activity of disc dog) in a video is useful for the prediction of the *Disc dog* action class. Recognizing *Lemon* and *Measure cup* is indicative for the anticipation of *Making lemonade*.

5. Conclusion

We have described a new model for future action anticipation. The main motivating idea for our method is to model action anticipation as a fusion of two complementary modules. The predictive approach is a purely anticipatory model. It aims at directly predicting future action given the present. On the other hand, the transitional model is first constrained to recognize what is currently seen and then uses this output to anticipate future actions. Our approach achieves state-of-the-art action anticipation performances on the EPIC-KITCHENS [6] and Breakfast [19] datasets.

	Action				Verb				Noun			
	A@1	A@5	P	R	A@1	A@5	P	R	A@1	A@5	P	R
Damen *et al*. (TSN Fusion) [6]	1.7	9.1	1.0	0.9	25.4	68.3	13.0	5.7	9.8	27.2	5.1	5.6
Damen *et al*. (TSN Flow) [6]	1.8	8.2	1.1	0.9	25.6	67.6	10.8	6.3	8.4	24.6	5.0	4.7
Damen *et al*. (TSN RGB) [6]	2.4	9.6	0.9	1.2	25.3	68.3	7.6	6.1	10.4	29.5	8.8	6.7
DMI-UNICT	**7.3**	18.8	**2.5**	**4.0**	27.2	69.3	**13.6**	9.2	12.4	30.7	**8.7**	8.9
Ours (Predictive)	6.1	18.0	1.6	2.9	27.5	**71.1**	12.3	**8.4**	10.8	30.6	8.6	8.7
Ours (Predictive + Transitional)	7.2	**19.3**	2.2	3.4	**28.4**	70.0	11.6	7.8	**12.4**	**32.2**	8.4	**9.9**

Table 4: **EPIC-KITCHENS results on hold-out unseen test set S2**. The official ranking is based on the action top 1 accuracy score (A@1). A@1: top-1 accuracy, A@5: top-5 accuracy, P: precision, R: recall. Challenge website details: https://competitions.codalab.org/competitions/20071. Note that our best model was submitted under the anonymous nickname **masterchef**.

Model	Accuracy
Random baseline	2.1
Vondrick *et al*. [43]	8.1
Abu Farha *et al*. (CNN) [7]	27.0
Abu Farha *et al*. (RNN) [7]	30.1
Ours (Transitional (AR))	23.9
Ours (Predictive)	31.9
Ours (Predictive + Transitional (AR))	**32.3**
Ours (Transitional (AR with GT))	43.0

Table 5: **Comparison to state-of-the-art on the Breakfast**. We report anticipation accuracy averaged over all of the test splits of Breakfast dataset[19] and use $T = 1$ *sec.*

Action to anticipate	Top-3 visual attributes
Applying sunscreen	Sunscreen, Lotion, Swimming trunk
Bull fighting	Ox, Bulldozing, Bullring
Camel ride	Arabian camel, Crane, Riding scooter
Disc dog	Border collie, Collie, Borzoi
Drinking coffee	Hamper, Coffee mug, Espresso
Making an omelette	Cooking egg, Wok, Shaking head
Making lemonade	Lemon, Measure cup, Pitch
Playing ice hockey	Hockey arena, Hokey stop, Teapot
Preparing pasta	Guacamole, Carbonara, Frying pan
Preparing salad	Wok, Head cabbage, Winking
Raking leaves	Hay, Sweeping floor, Rapeseed
Using parallel bars	Parallel bars, High jump, Coral fungus

Table 6: **Top-3 attributes that indicative of actions**. Top-3 visual attributes activations for the anticipation of some action class from the ActivityNet 200 dataset.

Acknowledgment. The project was partially supported by the Louis Vuitton - ENS Chair on Artificial Intelligence, the ERC grant LEAP (No. 336845), the CIFAR Learning in Machines&Brains program, and the European Regional Development Fund under the project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000468).

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016. 2

[2] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson. Encouraging lstms to anticipate actions very early. In *ICCV*, 2017. 1, 2

[3] G. Bertasius and J. Shi. Using cross-model egosupervision to learn cooperative basketball intention. *arXiv preprint arXiv:1709.01630*, 2017. 2

[4] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 1, 2, 5, 6

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 2, 5, 6

[6] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 1, 2, 4, 5, 6, 7

[7] Y. A. Farha, A. Richard, and J. Gall. When will you do what?-anticipating temporal occurrences of activities. In *CVPR*, 2018. 1, 2, 6, 7

[8] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 1, 2

[9] P. Felsen, P. Agrawal, and J. Malik. What will happen next? forecasting player moves in sports videos. In *ICCV*, 2017. 2

[10] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. In *NIPS*, 2016. 2

[11] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell. Actionvlad: Learning spatio-temporal aggregation for action classification. In *CVPR*, 2017. 1, 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 5

[13] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell. Localizing moments in video with natural language. *ICCV*, 2017. 1

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

[15] M. Hoai and F. De la Torre. Max-margin early event detectors. 2014. 1, 2

[16] A. Jain, H. S. Koppula, B. Raghavan, S. Soh, and A. Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *ICCV*, 2015. 2

[17] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012. 2

[18] Y. Kong, Z. Tao, and Y. Fu. Deep sequential context networks for action prediction. In *CVPR*, 2017. 1, 2

[19] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014. 1, 2, 5, 6, 7

[20] T. Lan, T.-C. Chen, and S. Savarese. A hierarchical representation for future action prediction. In *ECCV*, 2014. 1, 2

[21] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1, 2

[22] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li. Learning from noisy labels with distillation. In *ICCV*, 2017.

[23] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 1, 2, 3

[24] P. Luc, N. Neverova, C. Couprie, J. Verbeek, and Y. Lecun. Predicting deeper into the future of semantic segmentation. In *ICCV*, 2017. 2

[25] S. Ma, L. Sigal, and S. Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *CVPR*, 2016. 1, 2

[26] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *CVPR*, 2017. 2

[27] T. Mahmud, M. Hasan, and A. K. Roy-Chowdhury. Joint prediction of activity labels and starting times in untrimmed videos. In *ICCV*, 2017. 1, 2

[28] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016. 2

[29] A. Miech, I. Laptev, and J. Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 1, 5

[30] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh. Action-conditional video prediction using deep networks in atari games. In *NIPS*, 2015. 2

[31] S. L. Pintea, J. C. van Gemert, and A. W. Smeulders. Déja vu. In *ECCV*, 2014. 2

[32] B. A. Plummer, M. Brown, and S. Lazebnik. Enhancing video summarization via vision-language embedding. In *CVPR*, 2017. 1

[33] N. Rhinehart and K. M. Kitani. First-person activity forecasting with online inverse reinforcement learning. In *ICCV*, 2017. 2

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 5

[35] M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *ICCV*, 2011. 1, 2

[36] Y. Shi, B. Fernando, and R. Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *ECCV*, 2018. 1, 2

[37] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang. Autoloc: Weaklysupervised temporal action localization in untrimmed videos. In *ECCV*, 2018. 1, 2

[38] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *ICLR*, pages 568–576, 2014. 1, 2

[39] K. K. Singh and Y. J. Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 1, 2

[40] T. Suzuki, H. Kataoka, Y. Aoki, and Y. Satoh. Anticipating traffic accidents with adaptive loss and large-scale incident db. In *CVPR*, 2018. 2

[41] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, 2018. 1, 2, 5

[42] G. Varol, I. Laptev, and C. Schmid. Long-term Temporal Convolutions for Action Recognition. *PAMI*, 2017. 1, 2

[43] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, 2016. 1, 2, 5, 6, 7

[44] C. Vondrick and A. Torralba. Generating the future with adversarial transformers. In *CVPR*, 2017. 2

[45] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 2

[46] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, 2013. 1, 2

[47] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 1, 2

[48] X. Wei, P. Lucey, S. Vidas, S. Morgan, and S. Sridharan. Forecasting events using an augmented hidden conditional random field. In *ACCV*, 2014. 2

[49] T.-Y. Wu, T.-A. Chien, C.-S. Chan, C.-W. Hu, and M. Sun. Anticipating daily intention using on-wrist motion triggered sensing. In *ICCV*, 2017. 2

[50] Y. Xu, Z. Piao, and S. Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *CVPR*, 2018. 2

[51] T. Xue, J. Wu, K. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, 2016. 2

[52] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. C. Niebles, and M. Sun. Agent-centric risk assessment: Accident anticipation and risky region localization. In *CVPR*, 2017. 2

[53] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. 2017. 1, 2

[54] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 5