VLMAH: Visual-Linguistic Modeling of Action History for Effective Action Anticipation

Victoria Manousaki*[1,2], Konstantinos Bacharidis[1,2], Konstantinos Papoutsakis[3], and Antonis Argyros[1,2]

[1]Computer Science Department, University of Crete
[2]Institute of Computer Science, FORTH
[3]Department of Management, Science & Technology, Hellenic Mediterranean University

Abstract

Although existing methods for action anticipation have shown considerably improved performance on the predictability of future events in videos, the way they exploit information related to past actions is constrained by time duration and encoding complexity. This paper addresses the task of action anticipation by taking into consideration the history of all executed actions throughout long, procedural activities. A novel approach noted as Visual-Linguistic Modeling of Action History (VLMAH) is proposed that fuses the immediate past in the form of visual features as well as the distant past based on a cost-effective form of linguistic constructs (semantic labels of the nouns, verbs, or actions). Our approach generates accurate near-future action predictions during procedural activities by leveraging information on the long- and short-term past. Extensive experimental evaluation was conducted on three challenging video datasets containing procedural activities, namely the Meccano, the Assembly-101, and the 50Salads. The results confirm that using long-term action history improves action anticipation and enhances the SOTA Top-1 accuracy.

1. Introduction

Anticipating future actions during an observed complex activity is a critical ability that enables humans to recognize intended goals and outcomes to proactively plan and engage in interactions with other humans and the environment in a timely, efficient, and safe manner. We accomplish this task naturally by perceiving visual information and learning from a few activities as well as based on self-experimentation; thus, it encompasses harnessing relevant kinematic and contextual knowledge rooted in perception, personal experience, and skills. These competencies are regarded as fundamental constituents of human intelligence.

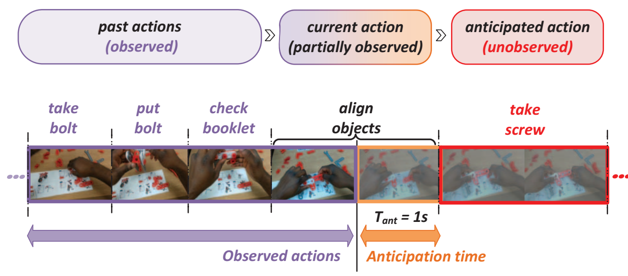

Figure 1. We consider the problem of action anticipation in untrimmed videos of procedural activities. At a certain moment in time (decision point), the proposed framework (VLMAH) anticipates the action (i.e., the unobserved action "take screw") that is most likely to be performed after some anticipation time T_{ant} (depicted with orange color). This is performed on the basis of the history of all past actions up to the decision point (depicted with purple) which is modeled by integrating visual input regarding the immediate past and a linguistic description of the distant past.

Deriving effective solutions for similar competencies is also beneficial to AI-enabled agents and robots that operate in industrial and domestic environments in a multitude of real-world applications [22]. In particular, the anticipation of near or long-term future actions can efficiently be used to advance autonomous navigation or driver-assistance systems, leverage the ability of industrial or home/socially assistive robots towards fluent human-robot collaboration and interaction, drive optimization of industrial workflows and enhance human safety through real-time hazard/anomaly identification to preemptively signal alerts and aids [38].

To enhance AI agents' capabilities, researchers have concentrated on video-based human understanding, yielding impressive outcomes in tasks like recognition, detection, and short- or long-term action prediction during ex-

*vmanous@ics.fort.gr (Corresponding author), kback@ics.forth.gr, kpapoutsakis@hmu.gr or papoutsa@ics.forth.gr, argyros@ics.forth.gr
https://projects.ics.forth.gr/cvrl/vlmah/

tended activities [22]. Among these, action anticipation stands out, involving forecasting upcoming action labels based on partial ongoing action observation and recent action history [42, 7], as depicted in Figure 1. The ability to use recent action history is crucial for proposing potential actions at the decision point T_{ant} before the expected start time of the next action. This anticipation time captures valuable insights and the sequence of actions leading to the anticipated one. We identify the following questions towards this challenging task, which effective solutions have to deal with by assessing the best trade-off between the complexity of spatiotemporal visual feature modeling and the accuracy performance of action anticipation:

- How much of the action history should be considered to accurately predict future actions during complex activities?
- What is the most efficient way to model the temporal ordering of action history (past actions)?
- What information modalities could enhance action anticipation accuracy?

We tackle the challenge of anticipating actions within instructional activities by merging visual and linguistic data from ongoing actions. This encompasses recent and distant history, vital for predicting the future. While visual features offer rich information, they are resource-intensive for storage and computation. In contrast, language-based action descriptions are less detailed but more storage and processing efficient. Our approach balances these aspects by integrating high-cost visual features for recent events and low-cost language features for remote ones.

We explore action anticipation in the context of procedural activities, where variations of the temporal ordering of actions are usually more constrained. Based on that, it is not surprising that the majority of existing works [14, 15, 58, 37, 31, 46, 61, 43, 18] aspire to tackle this problem using video datasets [7, 8, 27, 53, 24, 47, 41, 5] containing procedural activities. For instance, EpicKitchens [8] is one of the largest and most popular video datasets, among others [27, 53, 24, 50, 62], deals with the task of action anticipation featuring videos of cooking activities. Another popular domain of instructional activities that regard complex assembly activities [47, 41, 5, 19, 25, 39] in the context of industrial and non-industrial scenarios.

In particular, we focus on videos of assembly activities using the Meccano [41] and the Assembly-101 [47] video datasets. Those two can are considered complementary with respect to the types of the target activities, as participants in the former are provided with specific instructions to accomplish the assembly process of a toy vehicle, whereas in the latter participants were free to disassemble a fixed toy vehicle and then to assemble it from its parts, following a less constrained process.

Our contributions can be summarized as follows:

- We propose the Visual-Linguistic Modeling of Action History (VLMAH) framework that combines short-term visual and longer-term lexical information of observed past actions to estimate the label of the near-future anticipated action.
- We show that the combination of cost-effective processing and integration of linguistic information along with visual information can greatly benefit prediction accuracy in various types of procedural activities.
- An extensive experimental evaluation was conducted with state-of-art results on three challenging datasets, Meccano [41], Assembly-101 [47] and 50-salads [53], for a large set of different experimental setups, and anticipation times. VLMAH improves the noun/verb/action predictions for the Meccano and Assembly-101 dataset while for the 50Salads dataset, our method is amongst the top performing.

2. Related Work

Action/Activity Recognition sets the thematic base upon which more fine-grained video understanding tasks, such as action detection, early action recognition, and action anticipation/prediction have been defined. In its most challenging form, it comprises the recognition of actions that involve human-object interactions, and action sets with high intra- and inter-class variability. With the advent of deep learning, video action recognition methods have become extremely efficient and effective in modeling short-range dependencies of actions with CNN-centered models [52, 6]. Moreover, the ability to model long-range dependencies of complex actions or long, composite activities has also been considerably improved using memorization layers, such as RNNs and their variants [60, 28], attention mechanisms [57, 2], and temporal frame dependency modeling at multiple time scales [11, 59].

The significant performance gains that have been witnessed in this field have also been fueled by the emergence of large-scale datasets [7, 36], that contain diverse action sets, viewing conditions (egocentric [7, 51, 16] or third-person [1, 26]) and videos in various contexts providing rich, multi-level annotation data and different information modalities. Such datasets enabled the design of multi-modal models that apart from appearance and motion, also exploit audio, gaze-related data, and most importantly language [20, 17]. In the concept of multi-modal action/activity modeling, the visual-linguistic fusion scheme is shown to be extremely effective at representing the variability of complex actions and activities. This mainly relies on the action-related knowledge that is extracted using the lexical description of the action sequence and transitions, which is presented in the form of a simple text label or rich transcription/captions per action [20]. This information can be further processed using text statistics [45]. Recently, deep learning language models [54, 56], have also been proposed acting as a complimentary information source to the visual representation, expressed with handcrafted [44, 45] or deep learned [32, 23, 3, 4] descriptors.

Action Anticipation/Forecasting is defined as the task of predicting the class(es) of one or more future actions for which no observations are available at the decision time [22, 26]. The tasks of prediction and anticipation have been well-explored for actions of various complexity that range from simple motion primitives of a single human action [34] or a human-object interaction [22, 18, 35] to long, composite, procedural or unconstrained activities [48, 33]. Anticipating the near-future actions is performed towards a limited set or even thousands of action categories [7, 47]. Forecasting of the next actions is performed at "anticipation time" in the video that can be set at variable time horizons ranging from short- to long-term predictions. Many existing approaches fix this important task parameter to 1 second prior to the start of the action of interest [30, 14], while others explore the predictability of actions for several seconds [40, 12, 31, 1, 21]. The problem was initially introduced in third-person videos [18, 1], but it has recently gained significant popularity in first-person (egocentric) videos [7, 16], too.

The prominent method of Furnari *et al*. [13] explored the problem of action anticipation using "rolling-unrolling" LSTMs in order to summarize past actions and make predictions for the verb, noun and action of the next segment for multiple anticipation times. In [49] a multi-scale temporal model is proposed so that the past actions are aggregated for the future actions to be iteratively predicted. This framework performs predictions for the next action with an anticipation time of 1 second and is also capable of performing dense anticipation considering a large number of anticipated action classes. Our work complies with both methodologies so that the verb, noun, and action predictions are made in the range of $[0.25, 2]$ seconds with a step of 0.25 seconds.

Natural language processing (NLP) initially gained popularity in the cooking domain since recipes naturally contain a large variety of texts with instructions on food preparation. These large texts of instructions have attracted the interest for predictions of the next unobserved steps of the recipe in natural language in the form of sentences. Sener *et al*. [50] created a hierarchical model for learning multi-step procedures of recipe datasets with text and visual context. Their zero-shot anticipation framework is able to transfer knowledge from large-scale text corpora to the visual domain for the prediction of coherent and plausible recipe instructions. The same authors improved their framework by integrating a temporal segment proposal method into the video encoder and additional losses at the recipe encoder to improve convergence [48]. By comparing to recipe generation networks they showed that this method can perform better even for unseen recipes and dishes. Contrary to methods [50, 48] that exploit text to provide information to the visual domain, Mahmud *et al*. [33] proposes a two-step approach where information on the visual spatiotemporal context of the observed actions and the linguistic labels of the anticipated actions along with scene context are incorporated for caption prediction. Text and/or captions of the observed actions are not utilized.

Our framework deviates from the aforementioned approaches that use NLP, as we do not focus on the prediction of captions/sentences of the near-future, still unobserved actions. Instead, we focus on using linguistic information complementary to the vision module [3, 4] for the encoding of the short- and long-term history of the observed past.

3. Proposed Approach

The proposed Visual-Linguistic Modeling of Action History framework noted as VLMAH, is shown in Figure 2. It features a two-stream three branch deep neural network design that comprises (a) a vision-based action anticipation sub-network, (b) an activity-level sub-network for temporal modeling based on natural language processing (NLP), and, (c) a vision-based action recognition sub-net. The action anticipation visual sub-net is able to estimate the next action given the visual representation of the current/ongoing action segment exploring the short- and long-term action dynamics. The action recognition sub-net exploits the same input to provide estimates for the current action class. Additionally, the NLP-driven activity-centric sub-net is responsible for the long-range temporal modeling of the relation of the current action to the previously observed actions to learn a stochastic model of the forthcoming action.

The last architecture stage combines the two representations (visual action anticipation sub-net & language modeling sub-net) to anticipate one of the following events, (a) the next action (fine-grained label), (b) the active object of the next segment (noun), or (c) the next motion motif (verb).

3.1. Visual Action Anticipation Module

Given an input sequence x_t of the action y_t of an activity video sample $X_i = \{x_1, ..., x_N\} \rightarrow Y$, the visual action-anticipation sub-net aims at learning the representation of the on-going action at a segment-wise level, that will enable the prediction of the forthcoming action y_{t+1}. To achieve this, the proposed module follows a multi-branch design that operates on an ensemble of different vision-driven representations of the entire scene or of the key to the action

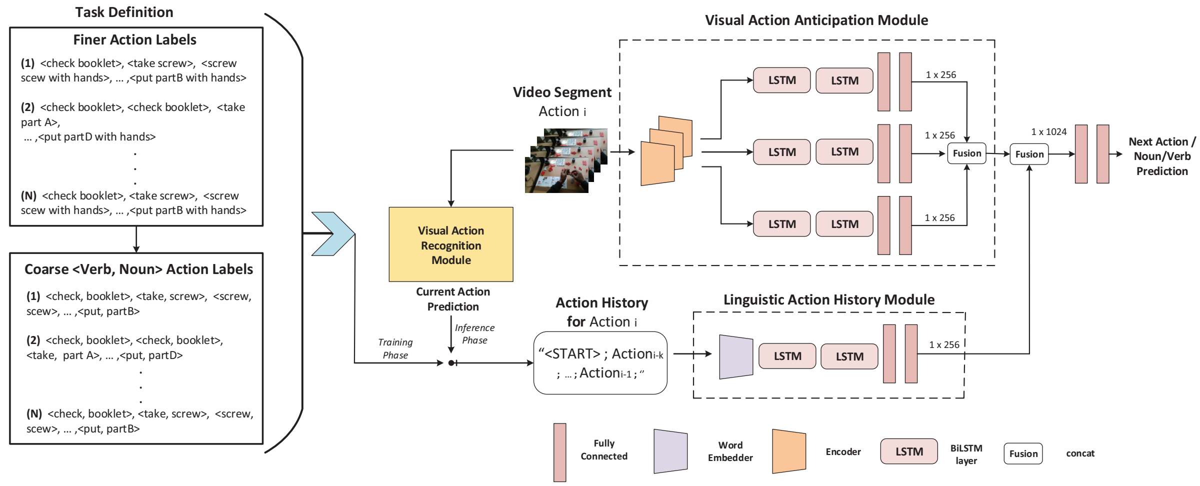

Figure 2. The proposed VLMAH architecture. The Visual Action Anticipation and the Linguistic Action History modules are presented. For the *Meccano* dataset, the encoders of the action module, generate Object, Hands, Gaze representations, whereas, for the *Assembly*-101 dataset, there is a single encoder network, TSM [29] while representations are split into 3 sub-sequences, as mentioned in Section 4.2. The detail level regarding the textual label descriptions is adaptable to the anticipation task at hand (action, motion motif (verb), or object (noun)). The final format also includes two special labels (START, END) that indicate the start and end of the action history sequence.

scene elements, such as the actor's body part regions or the appearance states of the active object.

On a technical basis, each branch of the proposed multi-branch design comprises a two-layer Bidirectional LSTM (BiLSTM) temporal encoder, followed by a Fully-Connected (FC) layer, that further encodes the representation into a $[1 \times 256]$ feature vector. Finally, the representations of all branches are fused via concatenation and forwarded to an FC layer that generates the final representation, which encodes the action segment into a $[1 \times 1024]$ feature vector. To form the inputs of this sub-net, we follow a sparse uniform sampling policy on the input sequence. Regarding the case of visual scene representation in the two datasets of interest, every single action of the action sequence that represents the activity has been encoded using a segment-wise temporal encoder network[1]. Therefore, it corresponds to the feature-based representation of a segment formed based on the adjacent frames. This formulation of the subnet's input enables the modeling of both short- and long-term appearance variations of the scene elements.

3.2. Linguistic Action History Module

We argue that the knowledge of the preceding action occurrences, noted as action history, is important for learning to estimate at a certain time in the video, the label of the next-anticipated action (*action forecasting/anticipation*) or of the active object in that action, as it provides efficient, discriminative features to opt among potential candidate targets. We address this issue using a compact textual description of the preceding actions, in the compact form of action labels, compared to captions that feature extensive textual descriptions of actions. The sentence-based textual description of the preceding actions is processed using the NLP sub-network that comprises a Word Embedding layer followed by the same layer set as the branches of the action-centric visual module. This representation forms a $[1 \times 256]$ feature vector, which is concatenated with the representation of the action-centric module. The combined representation is then forwarded to a set of two FC layers to provide estimations on the next action/object class.

Delving into this representation of the action history, we restructure each label (length, semantic complexity, part-of-speech element position (verb, noun, adverb)), in a specific lexical format depending on the task at hand (action, motion verb, or noun anticipation), to facilitate the learning process. Specifically, in the case of the verb (coarse motion motif) or noun (next-segment active-object) anticipation, we may have to deal with actions of a similar motion and object basis but of a different type of object upon which the action is performed. For example, consider the actions, *screw a screw with hands* and *screw a screw with screwdriver*. When asked to predict the key object(s)[2] of the next anticipated action, the action history module should maintain the key objects of the preceding action segments, and therefore the knowledge that the tool-medium is of no importance in this coarser anticipation problem. A similar convention is also considered for the task of predicting only the coarse motion motif label for the next action.

[1] For example, in Assembly-101 each action instance has been encoded using the effective Temporal Shift Module (TSM) [29]

[2] As key objects we refer to objects that affect the outcome of the activity, e.g. in a toy assembly activity on the parts that can alter the result.

Under this premise, for the tasks of verb/noun next-segment prediction we restructure the available lexical information/labels of actions by discarding parts of the labels that refer to the usage of extra tools (annotated as nouns) to implement the corresponding action, i.e. the action labels are restructured to follow the format *action verb + noun*. In fact, this meta-processing of action labels that allow for a decoupled prediction of the next action verb or next action object(noun), is a common practice followed by the recent datasets targeting isolated motion motif or next-segment object prediction (e.g. Assembly-101 [47]). If such an action label format is available for the dataset in question, this label restructuring is skipped. The gain from such lexical decomposition is that the prediction task becomes simpler since the number of classes decreases, due to the fact that labels sharing the same action verb or action object (noun) are being merged, which allows for more samples to be associated with the specific motion motif or object state. Finally, in the case of the next action prediction (entire action context), we do not restructure the initial labels since the entire context of the preceding action labels is required to disambiguate between actions that share the same motion and object characteristics but differ on the execution medium.

3.3. Visual Action Recognition Module

The two aforementioned modules can be regarded as independent action anticipation models. In addition, a visual action recognition model is incorporated independently which during the inference stage operates on the same input sequence, denoted as x_t, as the action y_t. The purpose of this model is to provide estimates specifically for the current action y_t and fill the language-based action history.

Since the purpose of this model is to fill the action history, it remains independent from the action anticipation modules without any influence or connection, it can be trained separately and applied during the inference stage of the framework. In this work, instead of developing and training an action recognition module from scratch, we leverage the capabilities of state-of-the-art (SOTA) action recognition models that have been documented in the existing literature for each dataset. This approach is motivated by our objective to construct a visual-linguistic action anticipation framework, which can benefit from the advancements achieved by action recognition models specific to each dataset, thereby enhancing its overall performance.

4. Experimental Setup

We evaluate the proposed framework on three popular datasets of procedural activities. The main characteristics of the datasets are described in this section, such as the target activities, camera viewpoints, annotation data as well as multi-modal data and features provided (Section 4.1), followed by the evaluation protocols.

The experimental evaluation for the proposed framework follows a two-way narrative. Firstly, the population of the action history module involves simulating the prediction scores of a realistic action recognition model on a given dataset. This step aims to showcase the model's performance in relation to the latest advancements for each dataset. Subsequently, the complete potential of the model is presented by populating the action history module with past predictions obtained from an ideal visual action recognition model for each respective dataset. We should note that the realistic visual action recognizer performance follows the current SOTA action recognition scores reported for each examined dataset. Finally, we conduct experiments regarding different portions of the linguistic action history to assess the effect of the different action history sizes on the anticipation capabilities of the proposed framework.

4.1. Datasets

Meccano [41] is a multi-modal egocentric dataset created to study the interactions of humans and objects in industrial settings during instructional activities. Twenty different participants were requested to build a toy model of a motorbike. There exist 20 object classes, which include 16 classes that categorize 49 different toy components, 2 tool classes namely the screwdriver and the wrench, the instructions booklet, and a special class, noted as a partial model, for the under-construction toy object. Also, the dataset contains 12 verb classes and 61 action classes. In total, 20 videos are provided, 11 of which are used for training while the rest 9 videos are used for validation and testing.

The Meccano dataset provides gaze, object-centric features, and hands-centric features. The former type of features are computed based on the occurrences of the objects in each frame following the work of [12, 13]. Gaze features have been obtained by weighting the object-centric features with the distance between the center of objects bounding boxes and the gaze position in the image. The hand annotations of the dataset that contain the bounding boxes of both hands were used as hands-related features.

Assembly-101 [47] is a large-scale video dataset for the analysis and understanding of procedural activities regarding assembling and disassembling 101 ”take-apart” toy vehicles captured from multiple viewpoints. In total 362 unique data sequences were captured synchronously by 4 egocentric and 8 static cameras and annotated with more than 100K coarse and 1M fine-grained action segments, targeting the challenging tasks of action recognition, action anticipation, temporal action segmentation, and mistake detection. Participants were instructed to disassemble and then assemble a toy vehicle without any instructions, which enhances the variability of the temporal ordering of actions performed by the participants during the procedural activities. A set of 90 object classes is considered that includes 5 tools together with the "hand". Also, 24 verbs are included along with the object classes form 1380 fine-grained action classes. A 60% of the available videos is used for training, while the rest 15% and 25% are utilized for validation and testing, respectively. Of the 101 toys, 25 of them are shared between all splits which sets the dataset even more challenging. For the RGB input, 2048-D frame-wise features are calculated using TSM [29] with an 8-frame input.

50Salads [53] is a multi-modal third-person instructional dataset of cooking-related activities. Twenty-five different participants prepared a set two mixed salads. The dataset provides RGB videos, depth maps, accelerometer data, and high- to low-level activity annotations. The dataset consists of 17 action classes. We report top-1 accuracy averaged over the 5 pre-defined splits following the work of [42].

4.2. Training, Testing & Input Configurations

As noted in Section 3, the structure of the action-centered temporal modeling sub-net follows a three-branch design, that acquires three vision-centered input sequences.

For the Mecanno dataset [41], input refers to the available feature representations for a) *Gaze*, b) *Objects*, and c) *Hands*. For the Assembly-101 [47], the available TSM [29] features for the RGB videos are utilized, which refer to frame-wise $[1 \times 2048]$ feature vectors. We restructure this representation to fit in the action-centric visual anticipation sub-net, as follows: a) split feature vectors into a set of two $[1 \times 1024]$ feature vectors to drive input to the first two branches and b) uniform sub-sampling is applied on the feature vector of the current frame of size $[1 \times 2048]$ into a $[1 \times 1024]$ and then calculate discrepancies between the sub-sampled feature representation of the previous frame to form the input feature vector for the third branch. For 50Salads [53] we utilized pre-extracted I3D features from [10, 42], which correspond to frame-wise $[1 \times 2048]$ feature vectors, which were restructured in the form described for the ones of the Assembly dataset.

Regarding the training configurations, the batch size was set to 4 for all datasets. The loss minimization is performed using the Adam optimizer, with a learning rate of 0.001. The input sequence length was set to 8 frames, while a random clip cropping sampling scheme was utilized. During the inference phase, we simulated the performance of the realistic visual action recognizer, by exploiting the SOTA performance of SlowFast [11] for Meccano, with 49.66% top1 accuracy, of TSM [29] for Assembly101 with 39.2% top1 accuracy, and, of Therbligs [9] for 50Salads with 76.5%.

5. Experimental Results

5.1. Action Anticipation

Predicting future actions is challenging, while modeling and performance greatly depend on the designated time horizon of the predictions. More specifically, predictions can be made at different decision points in time (timesteps) prior to the start of the next segment. In order to establish an extensive performance assessment of the proposed framework, we adopt the evaluation protocol reported in Furnari *et al*. [13] where predictions are made at 8 different anticipation timesteps before the start of the near-future anticipated action. Noted as τ_{ant}, the set of anticipation time refers to discrete values in the range of $[2s, 0.25s]$ for a timestep of $0.25s$. At the same time, the upper limit of this interval, that is $0.25s$ is closest to the start of the anticipated action.

Meccano: For the prediction of each action, the input to our framework regards information originating from the selected anticipation time point and runs backward, toward the start of the video (see Figure 1). As described in the previous sections, we exploit visual information related to the recent past (visual-action module) for modeling the short-term action history and the long-term past with the linguistic action history module. We report Top-1/Top-5 accuracy of the predicted action of the next segment, according to the [41]. In this work, the authors proposed the RULSTM framework [13] to anticipate the next action. We employ the publicly available code[3] of RULSTM for Meccano to replicate the experiments and also provide accurate results for the prediction of the noun and the verb of the next action-segment. We utilized a combination of information based on haze, object-centric and hand-centric features that are provided by [41], as those are the most discriminative features according to their experimental evaluation.

We evaluate the proposed VLMAH framework for action forecasting using different anticipation timesteps (see Table 1), and under the use of a realistic and an ideal (oracle) action predictor (denoted as VLMAH and VLMAH$_{GT}$ respectively) for past actions that populate the action history subnet. Under the use of a realistic visual action recognizer for past actions, our framework is compared to [41] which is the baseline and currently the SOTA method for the Meccano dataset. Our method outperforms the SOTA in Top-1 accuracy for the noun, verb, and action scenarios for almost every anticipation time, by a considerable margin. We present to have a slight decrease in performance in the Top-5 accuracy for the verb and action scenarios. This happens due to the impact of the action recognizer in the linguistic action history from which we draw information for making predictions. Our accuracy margin increases significantly from 4.1% up to 9.05% if we consider an ideal (oracle-like) visual action recognizer that feeds the linguistic action history module with the true past action classes. Any enhancement in action recognition accuracy is expected to similarly boost action anticipation, too.

[3]https://github.com/fpv-iplab/MECCANO

Top-1 / Top-5 Accuracy% @ different τ_{ant}									
		Timesteps							
Method		**2s**	**1.75s**	**1.5s**	**1.25s**	**1s**	**0.75s**	**0.5s**	**0.25s**
Meccano [41]		30.89/65.14	30.50/65.11	30.99/66.17	30.85/65.92	30.53/66.49	31.10/67.06	31.10/67.84	31.24/70.00
VLMAH		**33.12/77.85**	**32.12/77.78**	**31.48/78.49**	**32.33/80.41**	**31.25/76.30**	**32.17/82.39**	**34.07/78.58**	**38.34/79.19**
VMAH $_{GT}$	**Noun**	*15.91/72.58*	*27.63/69.46*	*25.37/65.83*	*28.93/73.29*	*26.21/70.31*	*25.08/71.73*	*28.83/69.81*	*29.50/70.88*
VLMAH $_{GT}$		*37.57/79.40*	*41.33/82.88*	*35.09/80.75*	*35.65/79.33*	*39.35/82.31*	*40.55/84.94*	*39.55/81.24*	*40.63/80.54*
Meccano [41]		36.06/ **93.19**	35.11/**93.01**	34.96/**92.98**	35.92/**93.19**	35.32/**93.38**	35.39/ **93.62**	34.75/ **93.76**	35.00/ **93.83**
VLMAH		**36.35**/93.00	**35.42**/92.33	**35.61**/91.31	**35.96**/92.88	**36.73**/91.08	**36.30**/90.62	**37.19**/91.14	**39.06**/90.93
VMAH $_{GT}$	**Verb**	*25.71/87.85*	*29.75/87.64*	*25.71/88.06*	*29.11/89.48*	*27.48/87.99*	*25.92/85.51*	*25.78/86.57*	*31.25/84.30*
VLMAH $_{GT}$		*40.76/91.40*	*41.26/93.39*	*40.83/92.61*	*43.39/92.96*	*39.91/91.69*	*40.98/93.18*	*43.67/92.68*	*43.55/91.65*
Meccano [41]		23.37/**54.65**	23.48/**55.99**	23.30/**56.56**	**23.97/57.73**	24.08/**58.23**	24.50/**59.96**	25.60/**61.31**	28.87/**63.40**
VLMAH		**24.75**/54.23	**24.35**/55.16	**24.22**/53.09	22.79/53.98	**28.90**/58.13	**25.29**/53.16	**26.47**/56.71	**29.12**/ 58.01
VMAH $_{GT}$	**Action**	*27.20/49.08*	*28.91/51.63*	*26.99/48.57*	*28.98/52.20*	*28.62/50.49*	*26.99/49.94*	*27.77/49.86*	*28.03/51.70*
VLMAH $_{GT}$		*34.73/67.75*	*36.86/69.53*	*35.01/67.18*	*34.30/69.24*	*35.15/68.25*	*33.59/67.89*	*34.65/66.90*	*33.09/65.98*

Table 1. Action anticipation accuracy for different timesteps (prior to the beginning of the next segment) for the **Meccano dataset**. *VLMAH* $_{GT}$ and *VMAH* $_{GT}$ represent the two variants of the proposed method when *ground truth annotations* are used as the linguistic action history. VLMAH makes use of the Linguistic Action History module while the action history is generated from the visual action recognition module. The comparison is between the [41] and the VLMAH methods.

Top-1/Top-5 Accuracy% @ $\tau_{ant} = 1s$			
Method	**Noun**	**Verb**	**Action**
TempAgg [47]	17.19 / 55.65	24.20 / 75.38	08.62 / 27.73
TempAgg [47]*	18.99 / 57.29	28.52 / 77.16	09.00 / 29.79
VLMAH	**27.70 / 54.37**	**42.17 / 82.52**	**14.18 / 30.95**
VMAH $_{GT}$	*22.68 / 55.32*	*40.59 / 85.11*	*13.14 33.98*
VLMAH $_{GT}$	*55.27 / 83.89*	*61.12 / 93.03*	*34.26 58.89*

Table 2. Top-1/Top-5 accuracy results of [47] and the VLMAH variants on the **Assembly-101 dataset** for anticipation time $\tau_{ant} = 1s$, with or without the use of the linguistic action history module. TempAgg$*$ denotes the single-task learning variant.

Top-1 Acc% @ $\tau_{ant} = 1s$	
Method	**Action**
DMR [55]	06.20
RNN [1]	30.10
CNN [1]	29.80
TempAgg [49]	40.70
AVT [14]	**48.00**
VLMAH	43.58
VLMAH $_{GT}$	*55.49*

Table 3. Top-1 accuracy results on the 50Salads dataset for the anticipation time $\tau_{ant} = 1s$.

Assembly101: In [47] that have also introduced the Assembly-101 dataset, action anticipation is performed at the fixed timestep $\tau_{ant} = 1s$. To assess action anticipation performance in [47], the TempAgg [49] method is used[4]. Both the VLMAH and the TempAgg methods are trained to generate predictions at anticipation time $\tau_{ant} = 1s$ that are evaluated using the Top-1 and Top-5 accuracy measures. Since the test split of the dataset is not yet available, we train and test both methods on the training and validation splits, respectively, using the egocentric viewpoint and data captured by the *e4* camera which yields the best results according to the experiments reported in [47]. Both the proposed VLMAH and the TempAgg methods have been trained/tested on data captured by this specific viewpoint.

Table 2 presents the accuracy results at $\tau_{ant} = 1s$. We provide two results for our framework. We compare our work with the state-of-art on Assembly-101 dataset, the TempAgg [49] framework. Our work is a single-task learning framework so for a fair comparison we test TempAgg [49] under two learning settings, a multi-task and a single-task. The single-task setting is denoted with $*$ in Table 2. The proposed approach outperforms state-of-the-art performance for the verb, noun, and action predictions by a large margin for this large and challenging dataset, even in the case that the linguistic action history module is not used. In particular, by using a realistic visual action recognizer to populate the action history module, an increase in accuracy of 13.65% for the verb prediction, 8.71% for the noun prediction, and 5.18% for the action prediction for $\tau_{ant} = 1s$ was reported. Similarly to *Meccano*, the use of an oracle-like visual action recognizer to verify/correct past estimates in the history module further increases the action anticipation performance of the proposed method. Even if we use only the visual information (VMAH), we outperform the TempAgg$*$ framework in general for a minimum of 4% up to 12%.

50-Salads: In Table 3 we present the accuracy scores at $\tau_{ant} = 1s$, and compare our proposed framework with recent works that tackle action anticipation in this dataset. We can observe that under the use of realistic action, recognizer to validate/correct the past action estimates stored in the action history module, our method is only surpassed by AVT [14] ($\approx 4\%$), with our proposed action anticipation method however, having a vastly lower number of trainable parameters (AVT: 378M, Ours: 10M), and ease in adapt-

[4]Code online at https://github.com/assembly-101

ing/incorporating the current action recognition advancements in each dataset.

5.2. How much history is enough?

In this study, we conducted ablation analyses to evaluate the performance of our proposed framework under various scenarios that pertain to the linguistic action history module's role and the required amount of linguistic action history to enhance the predictability power of the framework. Despite the fact that action history can obtain long-term information faster and with less cost compared to the visual features one question to be answered is "*how much history is enough?*". To answer this we evaluate our framework on the Assembly-101 dataset with different lengths of linguistic action history. From the previous sections, we have acquired the results of the evaluation of our framework with the full linguistic history of the observed actions[5]. In this experiment, we assess our framework by reducing the linguistic history to different percentages. The history percentages are in the range from 0% to 100%. Zero percent indicates the use of the VMAH$_{GT}$ framework while all the other percentages imply the use of the VLMAH$_{GT}$ framework with different percentages of action history. In this experiment, we use the VLMAH$_{GT}$ instead of VLMAH in order to assess the effect of the available size of action history in case no errors from the Visual Action Recognition module are present in the action history. As seen in Table 4, the results differ between the action and the verb/noun predictions considering different amounts of observed history.

Initially, all experiments were performed using 100% of the textual action history, which referred to a memorization capacity of 854 actions (slowest assembler). Our experiments show that, for the task of fine-grained action anticipation (full label), considering the entire linguistic history was the best strategy since it allowed us to disambiguate between cases of candidate actions that exhibited high similarity in their preceding action history.

In contrast, for the prediction of the coarse-grained verb and noun classes our experiments indicate that considering a more recent history is the best strategy. We observe that considering a larger percentage of the action history on these cases introduces noise that results in a considerable decrease in prediction accuracy, potentially due to similarities in the sequence of verb/noun transitions between different assembly scenarios. This is a valid assumption since, as stated in Section 3.2, in these tasks the initial action labels were restructured into a two-part-of-speech label (*verb+noun*). This way, we discarded the fine-grained context of the label that refers to the mediums (tools) utilized to perform the action. For example, in the case of the action label pair "*screw cabin with screwdriver*" and "*screw cabin with hands*", which are two different action classes, the restructuring operation merged the two classes into the action "*screw cabin*". We note that in Assembly-101 similar format is provided as annotation data.

[5]A full history refers to the number of actions the slowest assembler from the training set performed to complete the assembling task.

Top-1/Top-5 Accuracy% @ $\tau_{ant} = 1s$			
History	**Noun**	**Verb**	**Action**
0%	22.68 / 55.32	40.59 / 85.11	13.14 / 33.98
1%	56.98 / 83.35	62.33 / 92.78	28.49 / 53.69
12.5%	56.86 / 84.08	62.83 / 93.40	28.20 / 51.38
25%	53.86 / 83.03	62.92 / 93.06	28.96 / 53.15
50%	56.92 / 84.54	63.99 / 93.16	27.13 / 51.02
75%	56.19 / 84.53	63.20 / 93.21	29.83 / 53.75
100%	52.16 / 83.81	61.12 / 93.03	34.51 / 58.44

Table 4. The Top1 and Top5 accuracy scores achieved by the proposed framework using variable lengths of the linguistic action history on the **Assembly-101 dataset**. Zero percent (0%) is equivalent to the use of VMAH$_{GT}$ variant, while other action history percentage values refer to the use of the VLMAH$_{GT}$.

6. Conclusions and Future Work

This paper assessed the impact of a language-driven history-logging method on action anticipation. This mechanism complements visual action representation by memorizing prior actions. We explored its performance and resilience across diverse past action misclassification rates and the length of encoded action history in anticipation tasks (action, motion motif, object). Our experiments reveal the strategy's benefits, notably enhancing scores on tough video datasets showing procedural activities. Moreover, the proposed method proves robust even with limited memory and high misclassification rates. Future research will investigate the effects of incorporating the history of preceding actions on long-range action anticipation and examine the impact of the temporal positions of miss-classifications (e.g., short-term and long-term past) on action anticipation accuracy.

Acknowledgements

This research was (a) co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research" Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities, and (b) supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for HFRI Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment" (Project I.C.Humans, Number: 91) and under the "3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers" (Project InterLinK, Number: 7678).

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[3] Konstantinos Bacharidis and Antonis Argyros. Improving deep learning approaches for human activity recognition based on natural language processing of action labels. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[4] Konstantinos Bacharidis and Antonis Argyros. Cross-domain learning in deep har models via natural language processing on action labels. In *International Symposium on Visual Computing*, pages 347–361. Springer, 2022.

[5] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 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*, pages 6299–6308, 2017.

[7] 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.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23.

[9] Eadom Dessalene, Michael Maynord, Cornelia Fermüller, and Yiannis Aloimonos. Therbligs in action: Video understanding through motion primitives. In *CVPR*, 2023.

[10] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3575–3584, 2019.

[11] 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.

[12] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019.

[13] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020.

[14] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021.

[15] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022.

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[17] Jenhao Hsiao, Yikang Li, and Chiuman Ho. Language-guided multi-modal fusion for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3158–3162, 2021.

[18] De-An Huang and Kris M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 489–504, Cham, 2014. Springer International Publishing.

[19] Youngkyoon Jang, Brian Sullivan, Casimir Ludwig, Iain Gilchrist, Dima Damen, and Walterio Mayol-Cuevas. Epic-tent: An egocentric video dataset for camping tent assembly. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[20] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021.

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

[22] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[24] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[25] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised

action segmentation by joint representation learning and online clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20174–20185, 2022.

[26] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 689–704, Cham, 2014. Springer International Publishing.

[27] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018.

[28] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 166:41–50, 2018.

[29] 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*, pages 7083–7093, 2019.

[30] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 704–721. Springer, 2020.

[31] Tianshan Liu and Kin-Man Lam. A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13904–13913, 2022.

[32] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[33] Tahmida Mahmud, Mohammad Billah, Mahmudul Hasan, and Amit K Roy-Chowdhury. Prediction and description of near-future activities in video. *Computer Vision and Image Understanding*, 210:103230, 2021.

[34] Victoria Manousaki and Antonis Argyros. Segregational soft dynamic time warping and its application to action prediction. In *International Conference on Computer Vision Theory and Applications (VISAPP 2022)*, pages 226–235, 2022.

[35] Victoria Manousaki, Konstantinos Papoutsakis, and Antonis Argyros. Graphing the future: Activity and next active object prediction using graph-based activity representations. In *International Symposium on Visual Computing*, pages 299–312. Springer, 2022.

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[37] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. Rethinking learning approaches for long-term action anticipation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 558–576. Springer, 2022.

[38] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, 106:104078, 2021.

[39] Konstantinos Papoutsakis, George Papadopoulos, Michail Maniadakis, Thodoris Papadopoulos, Manolis Lourakis, Maria Pateraki, and Iraklis Varlamis. Detection of physical strain and fatigue in industrial environments using visual and non-visual low-cost sensors. *Technologies*, 10(2), 2022.

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

[41] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1569–1578, 2021.

[42] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252, 2021.

[43] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Untrimmed action anticipation. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari, editors, *Image Analysis and Processing – ICIAP 2022*, pages 337–348, Cham, 2022. Springer International Publishing.

[44] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1194–1201. IEEE, 2012.

[45] Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 144–157. Springer, 2012.

[46] Debaditya Roy and Basura Fernando. Action anticipation using latent goal learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2745–2753, 2022.

[47] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[48] Fadime Sener, Rishabh Saraf, and Angela Yao. Transferring knowledge from text to video: Zero-shot anticipation for procedural actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

[50] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 862–871, 2019.

[51] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] 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.

[54] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[55] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 98–106, 2016.

[56] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[57] 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.

[58] Xinyu Xu, Yong-Lu Li, and Cewu Lu. Learning to anticipate future with dynamic context removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12734–12744, 2022.

[59] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020.

[60] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[61] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023.

[62] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.