OWL (Observe, Watch, Listen): Audiovisual Temporal Context for Localizing Actions in Egocentric Videos

Anonymous CVPR submission

Paper ID 26

Abstract

Egocentric videos capture sequences of human activities from a first-person perspective and can provide rich multi-modal signals. However, most current localization methods use third-person videos and only incorporate visual information. In this work, we take a deep look into the effectiveness of audiovisual context in detecting actions in egocentric videos and introduce a simple-yet-effective approach via Observing, Watching, and Listening (OWL). OWL leverages audiovisual information and context for egocentric Temporal Action Localization (TAL). We validate our approach in two large-scale datasets, EPIC-KITCHENS and HOMAGE. Extensive experiments demonstrate the relevance of the audiovisual temporal context. Namely, we boost the localization performance (mAP) over visual-only models by +2.23% and +3.35% in the above datasets.

1. Introduction

Egocentric videos capture the world using wearable cameras. Arguably, in these videos, localizing actions in time is top of mind [14]. In doing so, we could enable world-changing applications such as an episodic memory AI assistant for health monitoring. Localizing and recognizing human actions in egocentric videos imposes several challenges. Due to the capture nature, videos tend to be long and highly unconstrained w.r.t. the activities occurring on the stream. Given that the capture happens through a camera mounted on a person's head, challenging conditions such as undesired camera motions, occlusions, and poor-quality video make the problem of localizing and recognizing actions a complex task. Additionally, existing egocentric datasets, *e.g*. EPIC-KITCHENS [15], focus on localizing atomic actions that happen densely across long videos. Consequently, the performance of egocentric TAL lags far behind compared to that in the third-person setting [32]. Given such complexity, analyzing the relationships of actions and looking beyond visual cues is essential in an egocentric scene.

Despite its challenges, there are particular properties of the current egocentric datasets [15,36] to benefit TAL. Since the videos are *unedited* and *continuous*, the audio stream is synchronized with the visual stream, capturing the sounds and appearance of what is happening in the video at the moment. This is different from videos in traditional datasets that are curated from online video platforms like YouTube. In such datasets and due to the editing, audio might not correspond to the original sounds present in the scene. We argue that audio, in the egocentric video, plays an important role in assisting visual models to localize human actions. For example, looking at Fig. 1a, we notice a person reaching for something in a kitchen. Because of the camera view, we cannot see the object they are interacting with. Can we guess what exactly are they doing? By observing the lighting and the location (above the stove), we could imagine the interaction with the fan. But how can we discern if the fan was turned off or on? By hearing the sounds from the scene, you would not doubt that the person is *'turning off the extraction fan'*. The fan's distinctive humming noise and its disappearance indicate the action happening and its precise temporal endpoints.

Using temporal context has been proven to be effective for both action recognition and localization [11, 24, 33, 44, 47–49]. Temporal context might be even more informative in egocentric videos. For instance, at being unedited and continuous, actions unfold, with a more often than not, predictable sequence [21,22,24]. To illustrate how context can be helpful in localizing egocentric actions, we present a toy example in Fig. 1b. Looking at the sequence holistically, the scenario is clear: the recorder is preparing a glass of juice. If we look at each shot separately (imitating a neural network classifying a trimmed clip), we could probably struggle to recognize some actions. It is unclear that the box, which the recorder is grabbing from the fridge, then opening and closing, contains juice. When we see some orange liquid (and hear) pouring from it, we can guess it must be orange juice. The instances *'grab juice'* and *'pour juice'* are almost five seconds away, but still are informative

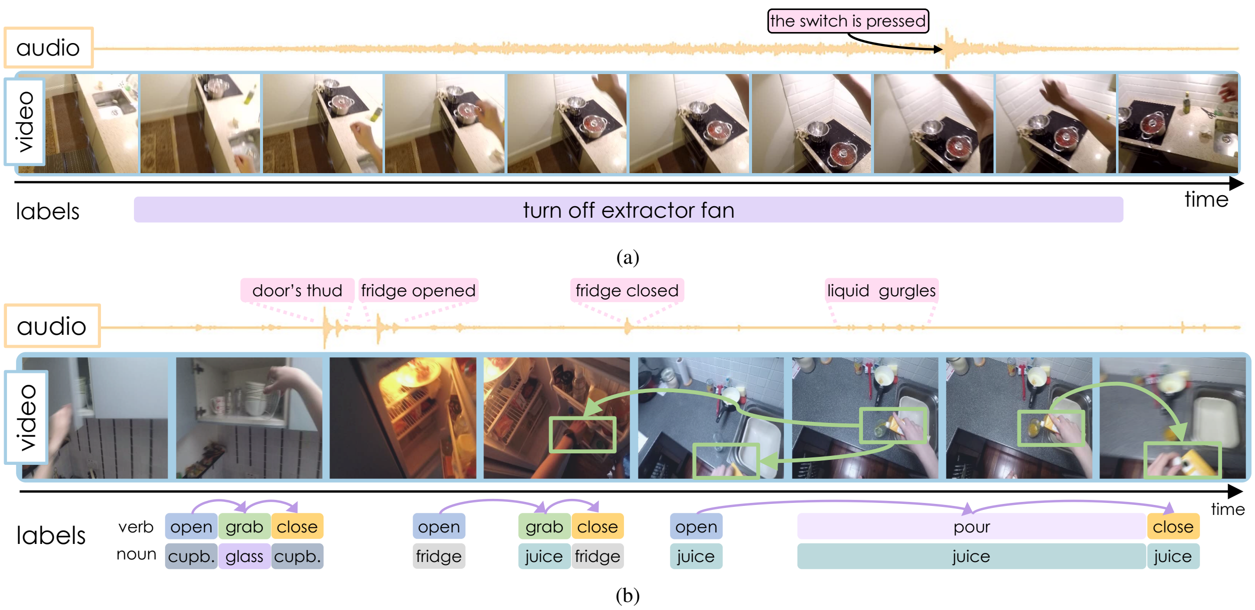

Figure 1. **Audiovisual temporal context is an important cue for the temporal localization of actions in egocentric unedited videos**. In video **(a)**, the action, *turning off the extractor fan*, is more evident when observing the interplay between audio and visual streams. The fan is invisible, but the interruption of the humming noise in the audio signal provides context to the movement of the hand in the visual domain. In video **(b)**, the recorder is preparing a glass of juice. The green drawn boxes spatially localize the juice box. Knowing the content of the box in action *pour juice* could help in predicting ambiguous actions *grab juice*, *open juice*, and *close juice* (green arrows). By following the violet arrows in the labels, we can see the pattern of how people interact with kitchen items (*e.g*. opening something, using it, then closing it).

to each other. Moreover, by leveraging context, we can decode the sequential patterns of actions in cooking activities. We argue that audiovisual context provides priors to better localize actions.

We propose **OWL (Observe, Watch, Listen)**, a simple-yet-effective transformer-based architecture that leverages audiovisual context to localize actions in egocentric videos. We do a methodical analysis to verify the importance of audiovisual context in egocentric videos. First, we study which components of the action localization pipeline would benefit from audio cues (Sec. 4, Tab. 1). Furthermore, we analyze what temporal neighborhood provides the richer context (Sec. 4, Tab. 3a). Finally, we analyze how visually occluded instances largely benefit from the context in egocentric videos (Sec. 4, Tab .6). OWL uses self-attention to encode context within each modality and cross-attention to capture relevant context across modalities. Our experiments on EPIC-KITCHENS [15], and HOMAGE [36] validate that OWL effectively encodes audiovisual context for egocentric TAL and significantly improves over proposed audiovisual baselines.

Contributions. The contributions of our work are three-fold: **(1)** We propose a transformer-based method for egocentric action localization by effectively fusing audiovisual context (Sec. 3). **(2)** We conduct extensive experiments on EPIC-KITCHENS and HOMAGE in Sec. 4.3, and achieve competitive results. **(3)** We conduct a thorough analysis that validates our hypothesis and findings about the audiovisual context for action localization in egocentric videos (Sec. 4.4).

2. Related Work

Audiovisual learning. Video and audio are common modality choices for a multi-modal learning scenario in video understanding. Deep learning facilitates audiovisual learning as it enables learning per-modality hierarchical representations [37], which are more optimal than designing hand-crafted features. Recent works provide us with more sophisticated solutions where the learned modality representations are being fused implicitly by the network and are optimized for the downstream task, such as [1, 17, 24, 25, 31, 43, 45]. While several works discussed the audiovisual scenario for the action recognition task [45], incorporating audio for TAL is not a widely researched area. [40] proposes a new task of audiovisual event localization that aims at predicting the event class from a 10-second clip. [4] studies multi-modal fusion approaches for audiovisual localization but ablates it on third-person datasets. Compared to them, we design our method for long, diverse egocentric videos. We are particularly motivated by [25], who emphasized the advantage of using egocentric unedited videos for applying audiovisual learning in action recognition. To the best of our knowledge, we are the first work that analyzes this advantage in egocentric TAL.

Temporal action localization (TAL). Given an untrimmed video, TAL models aim to detect the boundaries and classes of all actions happening inside the video. Recent work can be categorized into separate-stage and combined-stage methods. The separate-stage methods generate a set of class-agnostic proposals (generation) first and then use a separate classifier to assign an action class to each proposal [5, 8, 18, 27, 28, 47]. Most of the existing separate-stage methods focus on generating better proposals and rely on global video classification models and dataset statistics to classify them. Combined-stage solutions perform action localization in one unified pipeline by optimizing for both tasks simultaneously [29, 32, 46, 50]. In this paper, we follow the separate-stage approach.

Egocentric (unedited) videos. TAL has been extensively studied for third-person, and mostly edited videos (typically from consumer media platforms like YouTube and movies) [9, 23, 51]. The appearance of new large-scale egocentric datasets [15, 36] opened up a unique opportunity for researchers to study human actions in unedited videos. The annotations for action localization in most common (third-person) benchmarks are relatively sparse, with a low variation in assigned classes per video (ActivityNet [9] has on average 1.5 instances and 1.0 class per video, in THUMOS14 these numbers are 15.4 and 1.1, respectively). That makes it possible to condition the localized action class by gathering visual cues at the video level. This paradigm is not suitable for more dense and diverse datasets. For instance, EPIC-KITCHENS has, on average, 128.5 instances and 53.2 classes per video. That said, assigning proposals with a single video-level class would yield pretty poor localization results. To address the densely annotated videos on EPIC-KITCHENS, Damen *et al.* introduce a baseline separate-stage approach using BMN [27] proposals and SlowFast [20] classification. [32] proposes a combined-stage method (AGT) that leverages graph-based and transformer-based architectures to localize and classify actions jointly. Note that these approaches do not explicitly (or implicitly) model temporal context or leverage the egocentric audio streams. Our work lies in the separate-stage group; thus, to design OWL, we thoroughly investigate effective multi-modal and contextualized classifiers to assign each proposal an action class.

Temporal context in action localization. The importance of temporal context has been a long-standing aspect in action localization [2, 13, 35, 44, 47, 48]. Some works [47, 48] propose graph-based methods, where they define proposals and snippets as graph nodes and perform graph convolutions for the information exchange. Our approach is closer to recent work that leverages the Transformer architecture [30, 32, 39]. Due to the rising popularity of transformers for vision tasks [3, 10, 16], a few works [30, 32, 39] extended the transformer building blocks to the inner working of TAL as a way to infuse temporal context between proposals. In contrast to the prior art, our work considers the interplay of multiple modalities, visual and audio, while also modeling the surrounding context of an action. By putting *audiovisual context* at the forefront, architectural differences arise in comparison to existing transformer-based approaches.

3. Methodology

Given a sequence of video frames $V = \{I_t\}_{t=1}^{T}$, the task of TAL is to predict a set of segments $\Psi = \{\tau_n, s_n, y_n\}_{n=1}^{N}$ with start/end timestamps τ_n, confidence score s_n and action class labels y_n. In our work, we consider both the visual and audio modalities of the video sequence. We first encode either modality into snippet-level features $\mathbf{x} \in \mathbb{R}^{D\times L}$ [18, 47], where L is the number of encoded snippets and D is the channel dimension. The feature encoder usually adopts the pre-trained backbone of an action recognition model, such as [20, 42]. Our approach follows a **separate-stage pipeline**, where *Proposal generator* $\mathcal{G}$ generates class-agnostics proposals $\Psi_{\mathcal{G}} = \{\tau_n, s_n\}_{n=1}^{N}$, and then the *Proposal classifier* $\mathcal{C}$ assigns a class label y_n to them (including background class), as shown in Fig. 2a.

Observe, watch, and listen. We propose OWL (**O**bserve, **W**atch, **L**isten), a transformer-based model [41], to leverage multi-modal context in TAL. It uses an encoder composed of a self-attention module to encode the audio features and a decoder composed of a self-attention and a cross-attention modules to encode the visual features and to fuse both modalities (Fig. 2b). Besides **watching** the visual signal and **listening** to the audio signal, our OWL is also able to **observe** each proposal in the context of its neighbors' proposals. We model the visual and audio proposal-level features $\mathbf{z}^v$ and $\mathbf{z}^a$ as the input tokens for the transformer. We use the superscripts v and a for the visual and audio modalities, respectively.

Positional encodings. As transformer operations are permutation invariant, we use positional encodings to preserve the temporal relationship of the proposals. We encode the relative proposal start time and its absolute duration. The relative start time p_r incorporates the position of an action in the video and the temporal order of actions. By encoding the absolute duration p_d, we inject the temporal information that is lost after pooling. Specifically, $p_d = t_\text{e} - t_\text{s}$ and $p_r = \frac{t_\text{s}}{T}$, where t_e and t_e are the proposal's predicted start and end times, respectively. We pass p_r and p_d to a fully-connected (FC) layer to generate the positional encoding

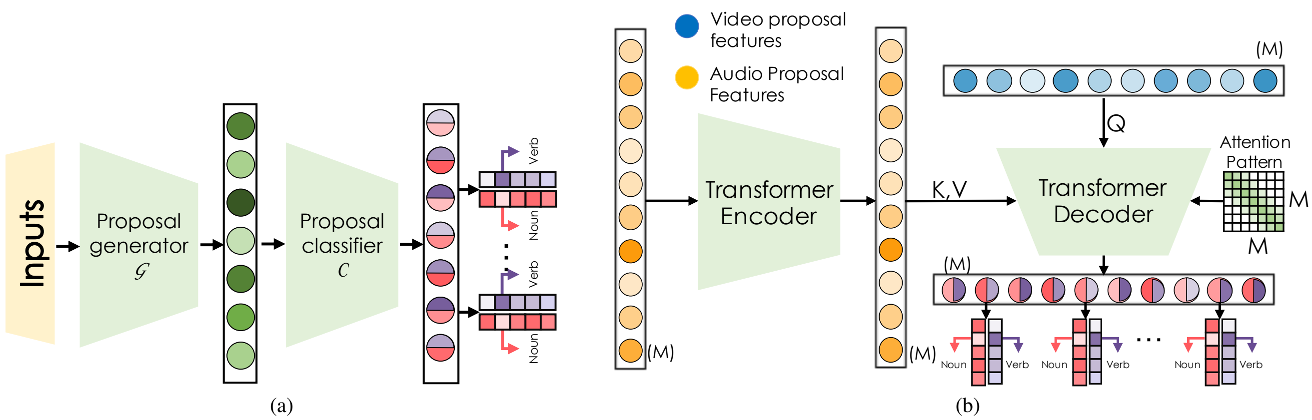

Figure 2. **(a) Separate-stage pipeline for TAL.** Given a sequence of snippet features, $\mathcal{G}$ produces class-agnostic action proposals with start/end timestamps. Then, $\mathcal{C}$ takes a set of proposal features and produces classification labels for each proposal. **(b) OWL:** We input the auditory sequence (yellow) into the encoder and the visual sequence (blue) into the decoder. K, V, and Q refer to the components of multi-head attention as in [41]. The encoder and decoder first perform self-attention to enrich the intra-modal representations. Then, the decoder performs multi-head cross-attention. The amount of context W (the green band on the attention pattern), within which self-attention and cross-attention act, can be controlled by the attention mask of size $M \times M$. M is the input sequence size (number of proposals).

$\mathbf{p} \in \mathbb{R}^{D^e}$ [19]. $\mathbf{p}$ is concatenated to $\mathbf{z}^v$ and $\mathbf{z}^a$ and passed to the transformer encoder.

Intra-modal & inter-modal context. For each token of either modality, the self-attention module **observes its relevant intra-modal context**, correlating other proposals to enhance its feature representation. After self-attention, we obtain enhanced representations $\mathbf{z}_e^v$ and $\mathbf{z}_e^a$ for each proposal. The transformer decoder fuses both modalities. It contains a cross-attention module, which takes $\mathbf{z}_e^v$ and $\mathbf{z}_e^a$ as input tokens. The visual modality tokens are used as queries Q, and audio modality tokens are used as keys K and values V (Fig. 2b). Recall that attention mechanism transforms Q, K, V as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V. \quad (1)$$

Hereby, the audio features are linearly combined based on the similarities between video and audio proposal-level features. The resulting features are enriched by **observing the inter-modal context** from neighboring proposals. Theoretically, we can correlate all M proposals in a video, but to study how much context is needed, we restrict the self-attention and the cross-attention to attend only to the proposals within a temporal neighborhood W (inspired by [6]). As shown in Fig. 2b, each proposal can attend to only $\frac{W}{2}$ tokens from each side.

Training and inference. We generate classification scores based on the enriched proposal-level features produced by OWL. We train the classifier $\mathcal{C}$ using standard cross-entropy loss. For the datasets that have multiple annotations per instance (*e.g.* noun and verb in EPIC-KITCHENS), we optimize for both noun and verb classification with a joint loss. During inference, we multiply the scores of each noun and verb prediction pair to generate the action scores.

4. Experiments

4.1. Dataset

We evaluate our proposed method on two large-scale egocentric video datasets. **EPIC-KITCHENS** [15] contains 700 unscripted videos of people performing their daily kitchen routines. It has, on average, 129 annotated instances per video, which make it significantly harder to perform TAL compared to the established benchmarks [9, 23, 51]. Around 28% of actions overlap, and each annotated instance is composed of a verb and a noun pair describing an action performed with an object. Overall, there are 300 noun and 97 verb classes.

HOMAGE [36] is a multi-view action dataset with audiovisual synchronized video data containing a diverse set of daily activities. It has, on average, 15 instances per video, and 90% of the scenes in HOMAGE have the egocentric view. The action annotations for HOMAGE are not decomposed into nouns and verbs, as in EPIC-KITCHENS. Therefore, we adapt our model to directly provide predictions for each action class. We train our model for 446 (out of 453) classes, as we removed some videos due to issues with the metadata.

4.2. Implementation Details

Features. For EPIC-KITCHENS, we experiment with audiovisual TBN [25], and a combination of visual SlowFast [20] and auditory SlowFast [26] features. We observe that using SlowFast features shows superior performance

Table 1. **Showing how uni-modal and multi-modal inputs affect the performance on EPIC-KITCHENS, measured by the average mAP.** *A*, *V*, and *AV* - auditory, visual, and audiovisual inputs, respectively (e.g. $\mathcal{G}$-*V* x $\mathcal{C}$-*AV* means that we input video features to proposal generator and audiovisual to the proposal classifier). We report the results on the validation set. Using audiovisual inputs results in a performance boost compared to using uni-modal inputs for both proposal classification and generation stages.

	Noun			Verb			Action		
	$\mathcal{C}$-A	$\mathcal{C}$-V	$\mathcal{C}$-AV	$\mathcal{C}$-A	$\mathcal{C}$-V	$\mathcal{C}$-AV	$\mathcal{C}$-A	$\mathcal{C}$-V	$\mathcal{C}$-AV
$\mathcal{G}$-A	2.00	9.01	9.81	2.00	8.17	08.94	0.45	5.65	6.70
$\mathcal{G}$-V	1.60	10.64	12.48	1.76	10.59	**11.96**	0.59	7.06	7.66
$\mathcal{G}$-AV	2.03	11.22	**12.63**	2.10	10.01	11.47	0.71	7.69	**8.35**

than TBN. Thus, we report all experiments using SlowFast features. We provide TBN experiments in the **supplementary**. We extract features at 5 FPS for training the proposal generator, and we max-pool them temporally for the proposal classification part. SlowFast features have a dimensionality of D = 2304. For EPIC-KITCHENS, both backbones are pre-trained on the EPIC-KITCHENS recognition task. For HOMAGE, the auditory SlowFast is pre-trained on VGG-Sound [12], and the visual on EPIC-KITCHENS.

Proposal generation. We use BMN [27] as our $\mathcal{G}$. In [27], the input is rescaled to a fixed size before being fed to the network. Given that the datasets are dense and contain mostly atomic actions, we implemented the sliding window approach (similarly to [34]). We use the sliding window of size 256 and the stride of 128 (160 and 80, respectively, for HOMAGE). We show the increase in average recall when using the sliding windows compared to the rescaling of the inputs in Table 2. As mentioned in [50], feature rescaling is suboptimal for detecting short actions in long videos. This is particularly relevant for our work as EPIC-KITCHENS is annotated with many atomic instances, and a video duration can exceed one hour. We also show the ablation for the best window size in the **supplementary**. We find a simple concatenation of visual and audio features, followed by an FC layer, to be an effective strategy to fuse the modalities (early fusion). We apply Soft-NMS [7] as post-processing.

Features	AR (%)
Rescaling	54.91
Sliding window	**64.61**

Table 2. **Proposal Average Recall (AR) on EPIC-KITCHENS** for the proposals treating the input sequence with rescaling *vs.* using the sliding windows. As videos can vary in duration, their features can have different temporal dimensions. We investigate two types of input sequence treatment in the proposal generator: (1) rescaling the features to produce the input of a particular temporal size and (2) iterating over the features with a sliding window. We can see that the sliding window approach results in a 10% AR increase compared to rescaling.

Proposal classification. In OWL, both the transformer encoder and decoder have 1 layer and 8 attention heads with a hidden unit dimension of 512. We experiment using learned or fixed positional encodings and find that the learned encodings perform better. The dimensionality of positional encodings $D^e = 32$. We also provide baselines with various multi-modal fusion strategies in the **supplementary**. These baselines perform worse than OWL.

4.3. Quantitative Results

Audiovisual impact. Before incorporating context with OWL, we validate a simple baseline to verify the impact of the auditory signal on $\mathcal{G}$ and $\mathcal{C}$. Here, instead of using the transformer, we simply concatenate audiovisual inputs and use FC layer to encode the proposal feature (no context).

We demonstrate the performance for 9 combinations of inputs in Tab. 1: $\mathcal{G}$ with visual (V) and/or auditory (A) inputs followed by $\mathcal{C}$ with visual (V) and/or auditory (A) inputs. We find that the audiovisual classifier ($\mathcal{C}$-AV) achieves the best results for all tasks (noun, verb, action). Furthermore, the audiovisual generator ($\mathcal{G}$-AV) performs the best for nouns and actions. This finding validates our intuition that *audio is a complementary signal to the video for detecting egocentric actions for both localization and recognition*. Our hypothesis is that audio helps localize actions in situations where visual interactions are occluded (an obstacle, bad camera view), unclear (dark environments), or ambiguous and where the audio signal is strong enough and discriminative. We discuss these scenarios in Sec. 4.4. Note that our naive audiovisual baseline ($\mathcal{G}$-AV and $\mathcal{C}$-AV) achieves 8.35% and already improves the action mAP by 1.3% when compared to visual-only performance ($\mathcal{G}$-V and $\mathcal{C}$-V) of 7.06%. We will further refer to the visual-only model as VM.

Incorporating context. In Tab. 3a, we ablate on the attention window size W. We find that increasing the window size *does* improve the performance of our model, validating our theory that *the temporal context is useful for the proposal classification*. Specifically, for EPIC-KITCHENS, W = 32 (9.06%) and W = 64 (9.29%) give us the best action average mAP. Using a smaller window performs comparably to the audiovisual baseline. Enlarging the window

Table 3. **The effect of attention window size W** on in the transformer block described in Sec. 3. We report the performance on the validation set, measured by the average mAP (%). Each token on the attention pattern can attend to $\frac{W}{2}$ tokens from each side. The optimal context size is 64 for EPIC-KITCHENS and 4 for HOMAGE. As discussed in Sec. 4.4, the differences in the relevant context size can be explained by the annotation density of the datasets.

(a) EPIC-KITCHENS

W	0	4	16	32	64	128	256	512
Noun	12.52	13.33	13.32	13.22	**13.96**	13.89	13.23	12.64
Verb	11.86	11.60	11.39	12.15	11.67	**12.16**	11.64	11.53
Action	8.21	8.71	8.90	9.06	**9.29**	8.78	8.58	8.66

(b) Homage

W	0	2	4	5	6	7	8	9
Action	8.17	9.11	**9.59**	9.43	9.46	9.07	8.78	8.64

Table 4. **Action localization on HOMAGE.** We compare the performance of the visual-only model (VM) *vs*. OWL. OWL significantly improves over the proposed baseline.

Method	VM	OWL
Average mAP	6.16	9.51

further degrades the performance slightly, suggesting that temporally distant proposals become irrelevant. Similarly, for HOMAGE increasing W improved the performance and reached its peak of 9.59% with $W = 4$. Recall that EPIC-KITCHENS has, on average, ~8.6 times more instances per video. Overall, our findings are similar to the observation on the optimal temporal context in [24]. However, they measure the context window size in trimmed actions and OWL in proposals. As proposals are dense and noisy and can be classified as background, our optimal window size is larger.

Comparison with the state-of-the-art. We compare the performance of OWL on EPIC-KITCHENS with the existing methods in Tab. 5. OWL performs significantly better than the baselines [15, 32], and achieves 9.29% average mAP for the action class. For HOMAGE, to the best of our knowledge, we are the first work to explore it for TAL. As shown in Tab. 4, OWL achieves 9.51% average mAP, which is a decent performance for more diverse dataset activities and a good baseline score to encourage more contributions from future work. In addition, we compare OWL with VM to validate the effectiveness of our approach to incorporate audio. OWL significantly outperforms VM by 3.35% average mAP.

4.4. Performance Analysis and Qualitative Results

Visual occlusion analysis. We validate the hypothesis that *OWL helps to detect actions in visually occluded environments* by comparing the mAP of *more-occluded vs. less-occluded* instances. To define the occlusion level, we assume that the visual occlusion must happen in the place of hand-object interactions. We utilize the detected hand-object interactions in EPIC-KITCHENS [38]. We measure the percentage of occluded frames per action instance by considering a frame as occluded when the object's bounding box of the interaction is missing. Then, we divide the validation set into 3 disjoint partitions: *No occlusion*, *Low occlusion* ($< 8\%$), *High occlusion* ($> 8\%$). We empirically find that 8% of occluded frames balances the size of the three partitions. We then evaluate VM and OWL on these partitions and measure the improvement in performance. As shown in Tab. 6 both models achieve the lowest performance on *High occlusion* across all tasks. As we hypothesized, we achieve the highest performance boost when using OWL over VM on *High occlusion* instances (56.7% improvement on action task *vs*. only 18.0% with *Low occlusion* and 22.2% with *No occlusion*).

Per-class performance of OWL. In Fig. 3, we show a per-class performance comparison of OWL *vs*. VM on EPIC-KITCHENS. We plot the absolute performance improvement, measured by average precision (AP), for the noun and verb classes. We can observe that OWL performs better than VM for most verb and noun classes. We attribute the improvements to audio or context incorporation.

Audio. Verbs *drink*, *pour*, *crush* have distinctive sounds, and OWL performs better than VM in these classes. *Drink*, is an interesting case as the source of sound is very close to the camera microphone. As we expect, OWL improves by more than 10% in this class. Likewise, several nouns, such as *machine:washing*, *toaster*, *fridge*, *fan:extractor*, *microwave*, *kettle*, etc., are electronic appliances which usually have distinctive sounds when turned on/off and while operating.

Context. Several verbs, such as *transition* (used interchangeably with *move*, *walk in* in the dataset taxonomy), *open*, *put*, *close* are predicted better with OWL. We believe that the improvement in these verb classes can be attributed to context incorporation. As mentioned in Fig. 1, humans often do their routine kitchen activities following some patterns (logical order in human-object interactions). We also observed that in EPIC-KITCHENS annotations, packaged food is often annotated as if its content was revealed (*e.g.* a closed package is annotated as *peas*, or orange juice as in Fig. 1). Predicting the content of packaged food without context is challenging, even for humans. We find that numerous food classes are ambiguous when shown packaged, such as *salt*, *cereal*, *pasta*, *juice*, *meat*, *grape*, *nut*, *carrot*,

Table 5. **Action localization on EPIC-KITCHENS.** We measure mAP@tIoU for tIoU $\in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and the average mAP on the validation and test sets. For reporting results on the test set, we **do not use** validation set for training, compared to [15]. OWL significantly improves over the presented baseline methods.

Method	mAP (Val) for Noun classes @tIoU						mAP (Test) for Noun classes @tIoU					
	0.1	0.2	0.3	0.4	0.5	Avg.	0.1	0.2	0.3	0.4	0.5	Avg.
Damen *et al.* [15]	10.31	8.33	6.17	4.47	3.35	6.53	11.99	8.49	06.04	4.10	2.80	6.68
AGT [32]	11.63	9.33	7.05	6.57	3.89	7.70	-	-	-	-	-	-
OWL (ours)	**17.94**	**15.81**	**14.14**	**12.13**	**9.80**	**13.96**	**16.78**	**15.22**	**13.60**	**11.64**	**9.74**	**13.40**

(a) Noun

Method	mAP (Val) for Verb classes @tIoU						mAP (Test) for Verb classes @tIoU					
	0.1	0.2	0.3	0.4	0.5	Avg.	0.1	0.2	0.3	0.4	0.5	Avg.
Damen *et al.* [15]	10.83	9.84	8.43	7.11	5.58	8.36	11.10	9.40	7.44	5.69	4.09	7.54
AGT [32]	12.01	10.25	8.15	7.12	6.14	8.73	-	-	-	-	-	-
OWL (ours)	**14.48**	**13.05**	**11.82**	**10.25**	**8.73**	**11.67**	**16.78**	**15.43**	**14.01**	**12.73**	**11.24**	**14.04**

(b) Verb

Method	mAP (Val) for Action classes @tIoU						mAP (Test) for Action classes @tIoU					
	0.1	0.2	0.3	0.4	0.5	Avg.	0.1	0.2	0.3	0.4	0.5	Avg.
Damen *et al.* [15]	6.95	6.10	5.22	4.36	3.43	5.21	6.40	5.37	4.41	3.36	2.47	4.40
AGT [32]	7.78	6.92	5.53	4.22	3.86	5.66	-	-	-	-	-	-
OWL (ours)	**11.01**	**10.37**	**9.47**	**8.24**	**7.26**	**9.29**	**9.69**	**9.03**	**8.07**	**7.11**	**6.23**	**8.03**

(c) Action

Table 6. **Visual occlusion analysis.** As discussed in Sec.4.4, we breakdown the validation set of EPIC-KITCHENS into 3 partitions: *No occlusion*, *Low occlusion*, and *High occlusion*, based on the percentage of missing predictions of the hand-objects interactions [15, 38]. We then report the mAP of VM and OWL for each partition. Intuitively, when the object is out of the frame (occluded), the hand-object interactions are missing. OWL improves the performance across all partitions, especially in *High occlusion* subset.

	No occlusion			Low occlusion			High occlusion			Validation set		
	noun	verb	action	noun	verb	action	noun	verb	action	noun	verb	action
VM mAP	16.0	14.8	10.8	14.3	16.0	12.8	9.4	10.0	6.0	10.6	10.6	7.1
OWL mAP	19.4	16.4	13.2	17.7	20.0	15.1	12.9	14.5	9.4	14.0	11.7	9.3
Improvement in %	**21.3**	**10.8**	**22.2**	**23.8**	**25.0**	**18.0**	**37.2**	**45.0**	**56.7**	**31.2**	**10.2**	**31.6**
# instances	4879			2407			2382			9668		

etc. We attribute the improvement in these classes to the context incorporation in OWL.

Qualitative Results. Fig. 4 visualizes the localization results of OWL and compares them to the results of VM. As we can see, VM fails to predict *open juice* and *close juice* actions. However, OWL predicts them successfully. We believe that the presence of context helps the model to understand the sequential relationships of the actions (opening, performing an action, then closing). Furthermore, the localized actions are more temporally precise with OWL (*open fridge*, *pour juice*). Our intuition is that the fridge and pouring sounds help to localize the actions better. We present more qualitative results in the **supplementary**.

5. Limitations

The scope of this work is limited to audio-visual context for TAL in untrimmed unedited videos with a large number of action categories per video. We acknowledge that the audio and visual signals may have an interesting interplay in highly edited videos, *e.g.* those on YouTube, TikTok, and movies. However, the audio signal in the edited video might not be predominantly associated with the action. It might

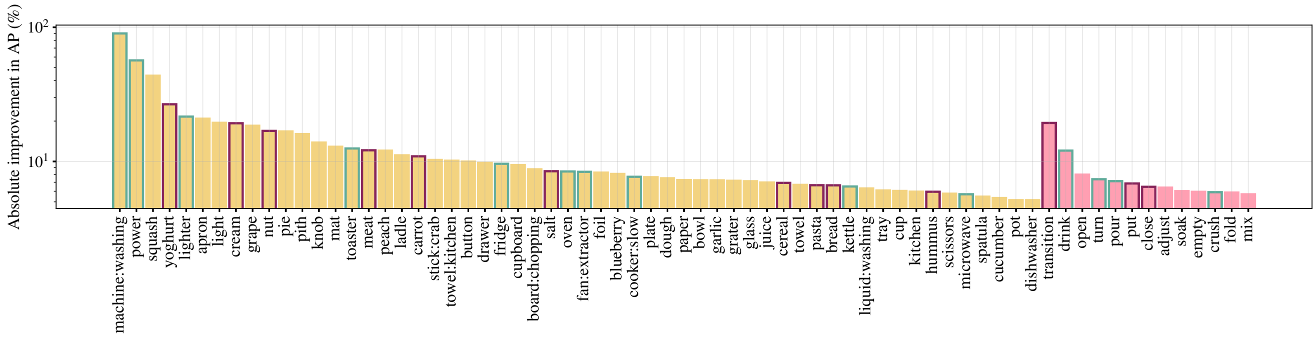

Figure 3. **Absolute per class performance improvement of OWL with respect to VM, measured by the average precision (AP, %).** To focus on the significant changes, we only visualize classes with the absolute difference in AP greater than 5%. We show the noun classes in orange and the verb classes in pink. As described in Sec. 4.4, we emphasize the bars of the classes, for which we attributed the improvement to the sound with green edges and to the context with red edges.

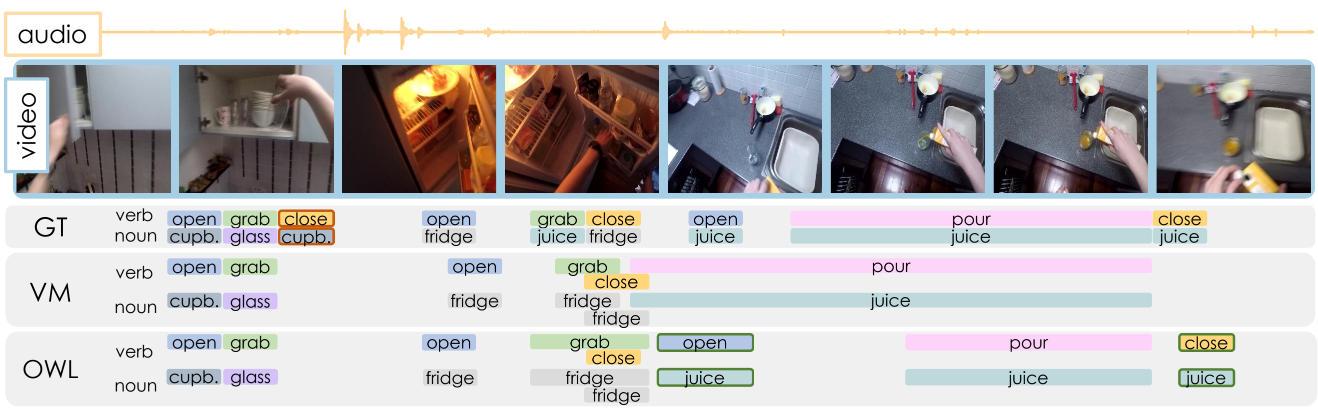

Figure 4. **Qualitative results for EPIC-KITCHENS.** The juice preparation scenario discussed in Sec. 1 (also described in Fig.1b). The ground truth (GT) annotations are compared with the predictions of VM and OWL. We highlight with red the GT actions that were missed by both VM and OWL (*close cupboard*) and with green the predictions where only either of them succeeds (OWL successfully predicts *open juice* and *close juice*). We can observe how OWL produces more temporally precise and complete outputs.

also be mixed with speech and music to evoke emotions in the viewers. Thus, we believe that it is healthy to explore the two lines of research independently, edited *vs*. unedited.

6. Conclusion

This work studies multi-modal TAL using egocentric unedited videos. The specific challenges of public egocentric video benchmarks (*e.g*., unedited footage, localization of actions out of frame, and a large number of action classes per video) invite us to rethink the inner workings of the TAL models. This work does so by means of two booster principles: multi-modality, with audio, and temporal continuity to complement the visual signal. We validate our hypothesis by experimenting with multiple audiovisual fusion approaches as well as context-aware pipelines. A technical contribution of our work is OWL, a transformer-based model that leverages both temporal context and modality fusion. By using OWL, we achieve competitive performance on EPIC-KITCHENS and set a strong baseline record on HOMAGE.

References

[1] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021. 2

[2] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. Action search: Spotting actions in videos and its application to temporal action localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 251–266, 2018. 3

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 3

[4] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. Hear me out: Fusional approaches for audio augmented temporal action localization. *arXiv preprint arXiv:2106.14118*, 2021. 2

[5] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–137, 2020. 3

[6] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4

[7] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 5

[8] Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem. Scc: Semantic context cascade for efficient action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1454–1463, 2017. 3

[9] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 3, 4

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[11] Alejandro Cartas, Petia Radeva, and Mariella Dimiccoli. Modeling long-term interactions to enhance action recognition. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10351–10358. IEEE, 2021. 1

[12] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 5

[13] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

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

[15] 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, 2021. 1, 2, 3, 4, 6, 7

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[17] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 2

[18] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3

[19] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Temporal localization of moments in video collections with natural language. *ArXiv*, abs/1907.12763, 2019. 4

[20] 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. 3, 4

[21] 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. 1

[22] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13505–13515, October 2021. 1

[23] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. Thumos challenge: Action recognition with a large number of classes. `http://crcv.ucf.edu/THUMOS14/`, 2014. 3, 4

[24] 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. 1, 2, 6

[25] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2, 4

[26] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 855–859. IEEE, 2021. 4

[27] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of The IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3, 5

[28] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[29] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11612–11619, 2020. 3

[30] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *arXiv preprint arXiv:2106.10271*, 2021. 3

[31] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *arXiv preprint arXiv:2107.00135*, 2021. 2

[32] Megha Nawhal and Greg Mori. Activity graph transformer for temporal action localization. *arXiv preprint arXiv:2101.08540*, 2021. 1, 3, 6, 7

[33] Yan Bin Ng and Basura Fernando. Human action sequence classification. *arXiv preprint arXiv:1910.02602*, 2019. 1

[34] Zhiwu Qing, Ziyuan Huang, Xiang Wang, Yutong Feng, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Changxin Gao, Marcelo H Ang Jr, and Nong Sang. A stronger baseline for ego-centric action detection. *arXiv preprint arXiv:2106.06942*, 2021. 5

[35] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 485–494, 2021. 3

[36] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 1, 2, 3, 4

[37] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017. 2

[38] 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 (CVPR)*, June 2020. 6, 7

[39] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. Class semantics-based attention for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13739–13748, 2021. 3

[40] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4

[42] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2016. 3

[43] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 2

[44] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019. 1, 3

[45] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

[46] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, page 5783–5792, 2017. 3

[47] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10156–10165, 2020. 1, 3

[48] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7094–7103, 2019. 1, 3

[49] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 1

[50] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13658–13667, 2021. 3, 5

[51] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8668–8678, 2019. 3, 4