# Weakly-Supervised Action Detection Guided by Audio Narration

Keren Ye*
Cruise
San Francisco, CA, USA
yekeren.cn@gmail.com

Adriana Kovashka
University of Pittsburgh
Pittsburgh, PA, USA
kovashka@cs.pitt.edu

## Abstract

*Videos are more well-organized curated data sources for visual concept learning than images. Unlike the 2-dimensional images which only involve the spatial information, the additional temporal dimension bridges and synchronizes multiple modalities. However, in most video detection benchmarks, these additional modalities are not fully utilized. For example, EPIC Kitchens is the largest dataset in first-person (egocentric) vision, yet it still relies on crowdsourced information to refine the action boundaries to provide instance-level action annotations.*

*We explored how to eliminate the expensive annotations in video detection data which provide refined boundaries. We propose a model to learn from the narration supervision and utilize multimodal features, including RGB, motion flow, and ambient sound. Our model learns to attend to the frames related to the narration label while suppressing the irrelevant frames from being used. Our experiments show that noisy audio narration suffices to learn a good action detection model, thus reducing annotation expenses.*

## 1. Introduction

Inexpensive and informative side information such as soundtracks and closed captions widely exist in videos. In addition, videos involve a temporal dimension, which synchronizes this side information with the video frames. Both the side information and the synchronization nature provide a good chance for self-learning. However, it is still challenging to achieve the goal of localizing specific actions using self-learning. On the one hand, which side information to use as supervision is unclear. On the other hand, the multiple modalities in videos such as RGB frames, motion features, and ambient sound need to be explored. Due to these complex factors, the idea of adopting the abundant side information to predict instance-level action detection results did not gather enough attention.

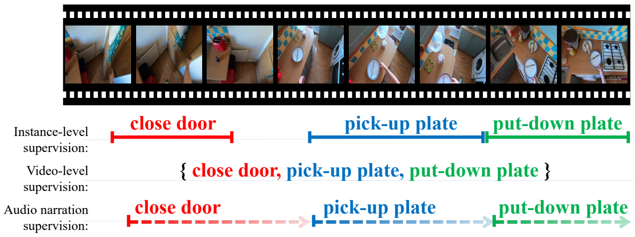
Instance-level supervision:
close door
pick-up plate
put-down plate
Video-level supervision:
{ close door, pick-up plate, put-down plate }
Audio narration supervision:
close door
pick-up plate
put-down plate

Figure 1. Demonstration of the instance-level, video-level, and audio narration supervisions. The audio narration supervision in the EPIC Kitchens dataset only includes an imprecise start time, while we use this cheap to annotate data to learn a video action detection model.

This paper will explore audio narrations in the untrimmed video action detection task. As the audio narrations roughly match the video frames, they provide a good signal for localizing actions in the temporal domain. We first distinguish the narration annotations from the instance-level or video-level annotations. We show in Fig. 1 an example video clip from the EPIC Kitchens dataset [12], as well as the different forms of supervision. The instance-level annotations are defined by triplets (start time, end time, action class). Models trained under the fully-supervised setting can use this form of supervision to generate temporal action proposals [32, 33], or perform action detection [14, 18, 34, 45, 46, 59, 61]. The major benefit of instance-level data is that the resulting models are usually boundary-sensitive, thus the foreground and background are clearly distinguished by the detection scores, resulting in high average precision.

However, fully annotating a video dataset with instance-level labels is time-consuming. Thus, methods [27, 39, 40, 42, 44, 53] focus on the weakly supervised action detection (WSAD), which only requires video-level labels. These methods assume the video to be a bag of actions and use multi-label cross-entropy loss to optimize. One disadvantage of WSAD is that it assumes only a few classes per video (e.g., $< 5$, see Tab. 1). Hence, it is not applicable

[*] Work done at University of Pittsburgh

| Dataset | Avg. video length (secs) | Avg. classes per video | Avg. actions per video |
|---|---|---|---|
| THUMOS 14 | 209 | 1.08 | 15.01 |
| EPIC Kitchens | 477 | 34.87 | 89.36 |

Table 1. Datasets information. Most WSAD methods use THUMOS 14 [21], in which there is only 1 action class per video. We explore single-timestamp audio narration annotations in EPIC Kichens [7].

in real cases. In an extreme scenario, a 2-hour untrimmed video may consist of all the action classes, thus the video-level label is too coarse to learn a good detector.

This paper explores the audio narration supervision in the EPIC Kitchens [12] dataset. In addition to the ambient soundtrack in the dataset, videos in EPIC Kitchens were narrated by the annotators. The resulting audio narration track is later transcribed into texts and then parsed into action classes (verb + noun) using a dependency parser, resulting in the forms we use (see Fig. 1 bottom). Our goal is to learn action detectors utilizing this form of cheap annotations, and we expect the performance to be comparable to the models learned using instance-level supervision.

The major challenge of using audio narrations is that the ***annotations are noisy***. Unlike the instance-level annotation: (1) narration annotations' start timestamps are not precise in that they may overlap with the previous action, and (2) the end time is unknown since the narrators provide no hints about it. One can assume the end time to be before the start time of the following action, but the frames in between are a gray area, and their membership is uncertain. Therefore, narrations provide a trade-off between accurate instance-level annotation and cheap and fast video-level annotation. One needs to model the uncertainty to use them.

To use the narrations to learn the action detection model, we first cut the untrimmed videos into clips using the single timestamp (start time) of the audio narration annotations (see Fig. 1). Thus, each clip can be treated as a mixture of actions, given that the boundaries are imprecise. Then, the association between the frames in the clip and the clip-level action class could be solved as a Multiple Instance Learning (MIL) problem. Compared to the common WSAD methods such as [39,40] which only distinguish between foreground and background (see statistics of THUMOS 14 in Tab. 1), **the background in our clip may involve other semantically meaningful actions**. So, we design a class-aware attention mechanism to assign higher scores to the frames in the clip that are more related to the narrated class. Meanwhile, we extract multimodal video features from RGB frames, motion flow, and ambient sound. We apply a simple early fusion architecture to the problem and ablate the contributions of each modality.

To summarize, our contributions are as follows:

- We propose to use the audio narrations to learn the video action detection model. To the best of our knowledge, this is a brand new task that has not been explored. In EPIC Kitchen tasks C1-C5, only C1-weakly is marginally related. However, C1-weakly is still different from ours because C1-weakly requires only to classify trimmed video at test time, while ours requires to localize actions in untrimmed videos.
- We provide a solution to the proposed task, in which we use a class-aware attention mechanism to rule out video frames that are not related to the narration label. Also, our solution considers multimodal features, including RGB frames, motion flow, and audio.
- We ablate our method on the EPIC Kitchens dataset and analyze the contributions of each model component and feature modality. The experiments provide insights for weakly supervised action detection methods in noisy untrimmed videos.

## 2. Related Work

**Weakly Supervised Learning in Images.** There is a large number of works aiming to harvest image models using weak supervision. Some use image-level labels [6,13,25,47,48,52,54] or captions [9,58,60] to learn object detection models. There are also works that learn semantic segmentation models [20,49,51] or scene graph generation models [57,62]. However, weakly supervised video models differ from all these prior image modeling approaches in that videos have the synchronized audio tracks, closed captions, and other information to be matched to the visual frames. Utilizing the additional synchronized modalities can potentially improve video models and is an important direction of learning video models using weak supervision.

**Weakly Supervised Learning in Videos.** Since videos naturally involve multiple modalities, many approaches use unsupervised or weakly supervised training to learn better video representations. For example, [5, 8, 38, 41] explore the cross-modal relations and leverage large amounts of unlabeled video for training. The basic idea behind this is that vision and sound are naturally synchronized so that models can utilize the synchronization as weak signals instead of ground truth labels. However, these methods are more often used in pre-training to improve the initialization of the visual-sound models. In comparison, our focus is on using additional modalities (e.g., audio narrations) to localize visual objects or actions in the temporal domain.

Also related is the co-localization or audio-visual correspondence [2–4, 16, 19, 43]. Similar to learning the joint representations, these works also rely on the synchronization of different modalities. However, they further learn to localize the sounds or visual objects given the information from other modalities. Our work differs from them in that

(1) these works did not quantify their results on detection tasks while only providing qualitative results; (2) our model requires no supervised signals at testing time.

**Weakly Supervised Video Detection Tasks.** In videos, there are various tasks of weakly supervised detection. For example, [27, 30, 31, 36, 37, 39, 53] only learn to detect the starting and ending time of particular actions, while entirely ignoring the spatial layouts of the instances. [56] use audio amplitude from the video to predict the occurrence of "climax" in an advertisement video. To track the spatio-temporal localization, methods such as [10, 55] rely on the video/image proposal frameworks such as [22–24, 50] which provide high-quality region proposals. Their approaches are counterparts to weakly supervised object detection in the image domain, with the key difference being in the types of proposals. Finally, there are also methods [28, 29, 35] attempting to only utilize cues from videos (e.g., motion, subtitle, tight boxes) to potentially benefit the training of image detectors.

We study how to learn action detection models (predicting starting/ending time and action labels) in videos. However, compared to fully-supervised methods [14, 18, 34, 45, 46, 59, 61], the supervised signals we used are the audio narrations, which are noisy in nature hence are much weaker than instance-level annotations. As for the weakly supervised action detection models [27, 39, 40, 42, 44, 53], their data in most cases only involves a single action per video. Thus video-level supervision satisfies their requirements. In comparison, our target task is a novel task, requiring non-trivial efforts to deal with the noisy annotations to improve the model's quality.

## 3. Approach

We first formulate the weakly supervised action detection (WSAD) task guided by audio narration, and overview the model training pipeline. Then, we introduce the details regarding the multimodal features in Sec. 3.1, discuss the design of the proposed class-aware attention in Sec. 3.2, and provide a post-processing algorithm which turns the frame-level into instance-level prediction (required by evaluation), in Sec. 3.3.

**Task formulation:** We conduct our experiments on the EPIC Kitchens dataset [12]. At training time, the video and paired $\{time_i, verb_i, noun_i\}_{i=1}^{N}$ as $N$ annotated actions are provided, where $time_i$ is the narration start time, and $verb_i$ and $noun_i$ are the narrated verb and noun classes respectively. The underlying assumption is that $time_i$ is not precise to represent the narration starting time since there may be overlap between consecutive actions. At test time, models have to predict four tuples $\{time\_s_i, time\_e_i, verb_i, noun_i\}$ given the video, where $time\_s_i, time\_e_i$ are the start and end time respectively.

**Training pipeline overview (Fig. 2):** Given a video and the paired audio narration annotation $\{time_i, verb_i, noun_i\}_{i=1}^{N}$, we first split the video into training clips. Given a specific action $(time_i, verb_i, noun_i), i \in \{1 \dots N\}$, we cut the video from $time_i$ to $time_{i+1}$, resulting in a video clip ($N$ clips in total) paired with $verb_i$ and $noun_i$. We denote the frames in the $i$-th clip as $\{f_{i,j}\}_{j=1}^{L_i}$ where $L_i$ is the total number of frames in the $i$-th video clip.

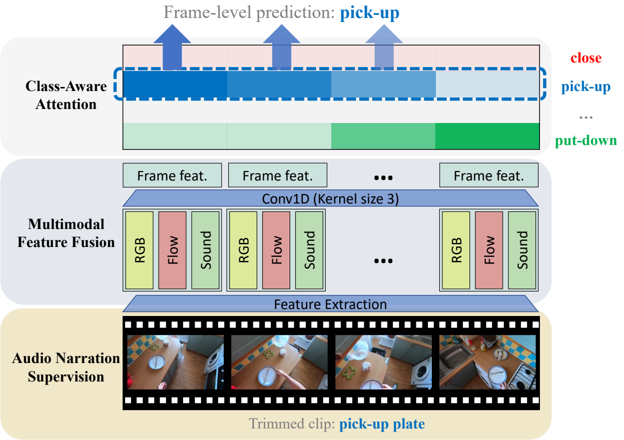

Figure 2. Model overview. We first cut the video into clips using the single timestamp denoted in the audio narration. Then, each video clip can be treated as a bag of a few actions. Next, we extract multimodal features and use early fusion to combine them (Sec. 3.1). We use a class-aware attention mechanism to produce the frame-level detection score (Sec. 3.2). Finally, we use a class-aware, intensity-sensitive post-processing (Sec. 3.3) to turn the frame-level into instance-level prediction (not shown), for evaluation purposes.

Then (Fig. 2 (middle)), we proceed with the feature extraction process, which will be explained in detail in Sec. 3.1. Briefly, we extract the visual CNN features of the RGB and flow frames and the semantic embedding of the ambient soundtrack. After feature extraction, we use early fusion to aggregate these multiple modalities.

Finally, we use an attention mechanism guided by audio narrations, to filter out irrelevant classes in the clip, given that the $i$-th clip should be all regarding the $verb_i$ and $noun_i$. In Fig. 2 (top), we show an example clip related to the "pick-up plate" action, and the figure demonstrates how to predict the verb "pick-up". Because of **the nature of the noisy narration supervision**, it is common that the last few frames in the clip mix with the next action (e.g., "put-down plate"). So given the clip level supervision "pick-up", how can we rule out the impacts of the next action "put-down"? Our solution is to predict different attention distributions for different verbs (in Fig. 2 (top), blue for "pick-up" and "green" for "put-down"). Then, within a clip, we use the attention distribution bound to the action class (e.g., verb "pick-up" is the supervision for the clip in

Fig. 2 (top)) to help with the predictions. The selected attention distribution will highlight the frames for predicting the clip-level action.

### 3.1. Multimodal Video Features

We consider features from the following sources. Though using them is common in video recognition, we are working on a novel task of learning action detection models from narration supervision, in which the contributions of RGB/flow/audio features are unclear.

- **RGB and flow frames.** We use the standard RGB and flow features provided in the EPIC Kitchens dataset, i.e., the 1024-D RGB and 1024-D Flow CNN features generated by a TSN model [15] pre-trained on [12].
- **Ambient sound.** Since the EPIC Kitchens dataset provides the soundtrack of the ambient audio, we also model them because the sound may imply some actions. We use VGGish [17] to produce a 128-D semantically meaningful embedding for every second. The VGGish method was first used in the AudioSet [17] classification task, and it was pre-trained on a large YouTube dataset (which later became YouTube-8M).

**Early fusion of the multimodal video features.** We linearly interpolate the ambient sound semantic embeddings to convert its sequence lengths to be the same as the RGB and flow features. We denote the concatenation of these multimodal features as the video frame feature $\{\boldsymbol{f}_{i,j}\}_{j=1}^{L_i}$, $\boldsymbol{f}_{i,j} \in \mathbb{R}^{2176\times 1}$ (RGB 1024-D, flow 1024-D, ambient sound 128-D, $L_i$ - number of frames). Let $\boldsymbol{F}_i = [\boldsymbol{f}_{i,1}\ \boldsymbol{f}_{i,2} \cdots \boldsymbol{f}_{i,L_i}]^T \in \mathbb{R}^{L_i \times 2176}$ be the video sequence feature, we apply a Conv1D layer (with kernel size 3, ReLu activation) to further extract the frame feature $\mathbf{F}_i = [\mathbf{f}_{i,1}\ \mathbf{f}_{i,2} \cdots \mathbf{f}_{i,L_i}]^T \in \mathbb{R}^{L_i \times d}$ ($d = 100$ is the number of neuron units):

$$\mathbf{F}_i = \text{Conv1D}(\boldsymbol{F}_i) \tag{1}$$

### 3.2. Class-Aware Attention for Weakly Supervised Action Detection

After getting the multimodal video features, we design a class-aware attention mechanism to localize the actions in the sequences. Our model selects relevant frames that best represent the action in the video clip and uses their aggregated features to represent it. Take Fig. 2 as an example, the verb class for the video clip is “pick-up” (i.e., clip-level label), so we use the embedding of “pick-up” to multiply (dot-product) each frame feature $\mathbf{f}_i$ to measure the frame-label similarity, resulting in a sequence of scores. After normalization, this score array represents the likelihood that the associated frames involve the action “pick-up”. We compute the weighted sum of the sequence features (weighed by the normalized scores). Then, we add a classification layer to predict the action and use cross-entropy loss to optimize.

Formally, we define action label embedding weights $\mathbf{W}^{(1)}_{verb} \in \mathbb{R}^{C_{verb}\times d}$, $\mathbf{W}^{(1)}_{noun} \in \mathbb{R}^{C_{noun}\times d}$ where $C_{verb}$ and $C_{noun}$ are the number of verb and noun classes, respectively. Since the verb detection and noun detection follow the same pipeline and only differ in the number of classes, we use $\mathbf{W}^{(1)} = \mathbf{W}^{(1)}_{verb}$ or $\mathbf{W}^{(1)}_{noun}$ as an abstract notation to denote either label embedding, $C = C_{verb}$ or $C_{noun}$ to denote the number of classes, and $c_i = verb_i$ or $noun_i$ to denote the clip-level label. Then, the following procedure applies to both verb and noun detection parallel tasks.

We first compute the dot-product between the label embedding and the frame feature, then, we use the sigmoid function to turn the score into a probability $\mathbf{A}'_i \in \mathbb{R}^{C\times L_i}$ (see Eq. 2; Fig. 2 (top) shows $\mathbf{A}'_i$ using the color matrix). Since we are aware of the class that is narrated in the video clip, we select the specific $c_i$-th row in $\mathbf{A}'_i$ ($c_i = verb_i$ or $noun_i$), resulting in $\mathbf{A}_i \in \mathbb{R}^{1\times L_i}$. This class-aware row selection process is shown in Fig. 2 (top) using the blue dashed box.

$$\mathbf{A}'_i = \text{sigmoid}(\mathbf{W}^{(1)}\ \mathbf{F}_i^T), \qquad \mathbf{A}_i = \mathbf{A}'_i[c,:] \tag{2}$$

Meanwhile, we use a fully connected layer $\mathbf{W}^{(2)} \in \mathbb{R}^{d\times C}$ to estimate the per-frame detection score $\mathbf{D}_i \in \mathbb{R}^{L_i\times C}$. In Eq. 3, $j \in \{1\cdots L_i\}$ denotes frame id and $k \in \{1\cdots C\}$ is the class index.

$$\mathbf{D}'_i = \mathbf{F}_i\ \mathbf{W}^{(2)}, \qquad \mathbf{D}_i[j,k] = \frac{\exp\left(\mathbf{D}'_i[j,k]\right)}{\sum_{k'=1}^{C}\exp\left(\mathbf{D}'_i[j,k']\right)} \tag{3}$$

Directly optimizing the per-frame detection score $\mathbf{D}_i = \mathbf{D}_{verb\ i}$ or $\mathbf{D}_{noun\ i}$ is hard since we only have the clip-level label $c_i = verb_i$ or $noun_i$. Thus, we apply the class-aware attention weighting $\mathbf{A}_i$ to aggregate frame-level information into $\bar{\mathbf{F}}_i \in \mathbb{R}^{1\times d}$ (Eq. 4), which is a clip-level feature. Then, the clip-level prediction is given by $\mathbf{P}_i \in \mathbb{R}^{C\times 1}$ (Eq. 5), which shares the $\mathbf{W}^{(2)}$ with Eq. 3.

$$\bar{\mathbf{F}}_i = \frac{\mathbf{A}_i\mathbf{F}_i}{\sum_{j=1}^{L_i}\mathbf{A}_i[j]} \tag{4}$$

$$\mathbf{P}'_i = (\bar{\mathbf{F}}_i\ \mathbf{W}^{(2)})^T, \qquad \mathbf{P}_i[k] = \frac{\exp\left(\mathbf{P}'_i[k]\right)}{\sum_{k'=1}^{C}\exp\left(\mathbf{P}'_i[k']\right)} \tag{5}$$

Finally, we use cross-entropy to optimize the model, where $\boldsymbol{y}_i$ is the one-hot representation of $c_i$ ($\boldsymbol{y}_i[k] = 1$ iff $k = c_i$).

$$L = -\sum_i \sum_{k=1}^{C} \boldsymbol{y}_i[k] \log \mathbf{P}_i[k] \tag{6}$$

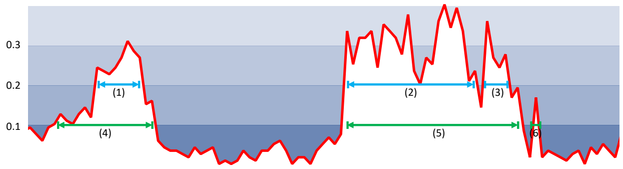

Figure 3. Intensity-sensitive post-processing. For each of the action classes, we use a set of thresholds (e.g., {0.1, 0.2}) and retrieve all segments (consecutive frames) that meet the different threshold conditions. The retrieval results are a set of action segments with different intensities. Next, we score each segment and apply Non-Maximum Suppression (NMS) to remove highly overlapped detections. We show the action clips detected using a threshold of 0.1 using green color and the clips detected by threshold 0.2 using blue. Assuming the IoU threshold of 0.6, segment (5) will be removed because it overlapped with (2).

### 3.3. Class-Aware Intensity-Sensitive Post Processing

To get the detections in the form of $\{time\_s_i, time\_e_i, verb_i, noun_i\}$ from the frame-level prediction $\mathbf{D}_i$ (Eq. 3), we use a class-aware intensity-sensitive post process. Specifically, we consider each action class separately. Given the detection score of a specific class (e.g., the $k$-th class in verb detection $\mathbf{D}_{verb\ i}[:, k]$), we first use different intensities (thresholds) to retrieve the segments, which are defined to be the longest sequence of consecutive frames that have detection scores past the threshold. The result is a set of potential action segments detected by different intensity scores (thresholds). We then assign a score to each segment, denoting the average detection intensity within the segment. In Fig. 3, we show the segments detected by threshold 0.1 and 0.2 using green and blue colors, respectively. In the next step, we use Non-Maximum Suppression (NMS) to remove highly overlapped (measured by IoU) detections and retain only those with higher intensity. Finally, we aggregate the NMS-ed detections from all action classes and sort them by intensity, resulting in our final detection results.

## 4. Experiments

We provide the details regarding our model in Sec. 4.1. Then, we provide experimental results in Sec. 4.2, including analysis regarding both the contributions of our model components and the benefits of multimodal features. To better understand our model, we also provide qualitative results in Sec. 4.3.

### 4.1. Implementation Details

Before training the detector, we extract the multimodal features offline. The CNN features of the RGB and flow frames are from [12], while we pre-processed the audio features. We use FFMpeg to extract audios from MP4 videos and feed the Mel spectrogram to the VGGish [17] model pre-trained on the large Youtube dataset (latter Youtube-8M) to produce semantic audio embeddings. After getting the above features, we interpolate the audio features to make them the same lengths as the RGB and flow features.

We concatenate the multimodal features as the model input and add a Conv1D layer (with $d = 100$ filters, kernel size 3, ReLu activation) to further finetune. During training, we use a dropout probability of 0.5 for the Conv1D layer, and a dropout probability of 0.5 for the learned attention ($\mathbf{A}_i$). We use the Tensorflow framework [1], Adam optimizer [26], a learning rate of 1e-5, and a batch size of 8 (8 clips). All models in our experimental sections are trained for 300K steps on the EPIC Kitchens dataset, using a validation set to pick the best model.

For the post-processing, we first apply uniform filtering (filter size 3) on each class's detection scores (e.g., $\mathbf{D}_{verb\ i}[:, k]$) to make the detection scores less fluctuating. Then, we vary the detection threshold from 0.01 to 0.4 to retrieve all segments and use NMS with an IoU threshold of 0.4 to remove highly overlapped segments.

### 4.2. Results on the EPIC Kitchens Dataset

**Metrics.** Although our training process does not rely on instance-level annotations, we can use the EPIC Kitchens' C2 task's (Action Detection) evaluation protocol, which measures the performance of the action detections. The protocol computes the average of the Average Precision (AP) values for each class, a.k.a. mean AP. A predicted segment is considered correct if its Intersection over Union (IoU) with a ground truth segment is greater than or equal to a given threshold (0.1 to 0.5). Besides the verb and noun detection, the EPIC Kitchens' C2 task also involves an action detection evaluation which requires the verb and noun detections to be correct at the same time.

**Contributions of Proposed Components.** We verify the effectiveness of the proposed model and compare it to the fully- and weakly-supervised action detection methods. All methods listed below use the same features.

- FUL. [11] is a fully supervised model trained by the EPIC Kitchens challenge organizer, using a two-stage approach to solve the action detection (action proposal [32] + action classification [14]).

- OUR FUL. is a one-stage fully supervised method trained by us, in which we predict the frame-level actions then post-process (Sec. 3.3). We treat OUR FUL. as a proper upper bound baseline in that all of our weakly supervised methods depend on similar frame-level prediction + post-processing.

- NARR. BAS. is the baseline method of using narration supervision. In NARR. BAS., we treat the single

| | Action Detection | | | | | | Verb Detection | | | | | | Noun Detection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. |
| Ful. [11] | 6.95 | 6.10 | 5.22 | 4.36 | 3.43 | 5.21 | 10.8 | 9.84 | 8.43 | 7.11 | 5.58 | 8.36 | 10.3 | 8.33 | 6.17 | 4.47 | 3.35 | 6.53 |
| Our Ful. | 6.40 | 5.69 | 4.59 | 3.34 | 2.39 | 4.48 | 12.9 | 11.4 | 9.04 | 6.62 | 5.03 | 9.00 | 11.4 | 9.61 | 7.17 | 4.70 | 2.98 | 7.17 |
| Narr. Bas. | 4.42 | 3.62 | 2.91 | 2.06 | 1.47 | 2.90 | 9.39 | 7.45 | 5.68 | 3.99 | 2.85 | 5.87 | 8.43 | **6.92** | **5.24** | **3.50** | **2.37** | **5.29** |
| Cls. Agno. | *4.57* | *3.78* | *3.10* | *2.28* | **1.70** | *3.09* | **10.0** | **8.53** | **7.03** | **4.79** | *3.40* | **6.75** | *8.49* | 6.82 | 4.96 | 3.22 | 2.04 | 5.11 |
| Ours | **4.68** | **4.01** | **3.27** | **2.33** | *1.65* | **3.19** | *9.64* | *7.96* | *6.31* | *4.70* | **3.56** | *6.43* | **8.51** | *6.88* | *5.09* | *3.36* | *2.25* | *5.22* |

Table 2. Contributions of proposed components. We show the Average Precision (%) at certain IoU thresholds (@0.1-@0.5) and the mean Average Precision (Avg.). Higher numbers are better. The best weakly supervised model learned using narration annotations is shown in **bold** and the second best is in *italic*.

| | Action Detection | | | | | | Verb Detection | | | | | | Noun Detection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. | @0.1 | @0.2 | @0.3 | @0.4 | @0.5 | Avg. |
| RGB | 4.49 | 3.76 | 2.94 | 2.25 | **1.67** | 3.02 | 8.72 | 7.07 | 5.43 | 4.38 | 3.15 | 5.75 | **8.70** | **7.13** | **5.29** | **3.71** | **2.56** | **5.48** |
| Flow | 2.32 | 1.98 | 1.47 | 1.10 | 0.84 | 1.54 | 6.59 | 5.58 | 4.29 | 2.95 | 2.10 | 4.30 | 4.33 | 3.47 | 2.49 | 1.68 | 1.11 | 2.61 |
| Audio | 0.34 | 0.27 | 0.23 | 0.09 | 0.05 | 0.20 | 1.71 | 1.37 | 1.07 | 0.61 | 0.39 | 1.03 | 0.94 | 0.68 | 0.51 | 0.23 | 0.16 | 0.50 |
| All | **4.68** | **4.01** | **3.27** | **2.33** | 1.65 | **3.19** | **9.64** | **7.96** | **6.31** | **4.70** | **3.56** | **6.43** | 8.51 | 6.88 | 5.09 | 3.36 | 2.25 | 5.22 |

Table 3. Contributions of multimodal features. We show the Average Precision (%) at certain IoU thresholds (@0.1-@0.5) and the mean Average Precision (Avg.). Higher numbers are better. The best model is shown in **bold**.

timestamp in the narration annotations as the boundaries and use the cut result as instance-level annotations to directly train a fully supervised model.

- Cls. Agno. is an alternative method, in which we use a class-agnostic attention instead of class-aware attention (Sec. 3.2).

Tab. 2 shows the results. We found Our Ful., though a one-stage method, is very competitive to Ful. [11] (action detection mAP 4.48% v.s. 5.21%). The only weakness is that it is not that good at boundary refinement. Hence its verb and noun detection AP@0.1,0.2,0.3 are higher, but its AP@0.4,0.5 are lower. Then, Narr. Bas., which uses the same fully supervised method (but changes to use the narration supervision), inevitably hurts the action detection performance (action mAP 2.90% v.s. 4.48%). This performance drop is due to the unclear boundary definition. We conclude that our method with uncertainty modeling (class-aware attention) helps to improve the use of narration supervision (action mAP 3.19% v.s. 2.90%). Also, we show that our modeling of class-aware attention is better than the alternative of class-agnostic attention (action mAP 3.19% v.s. 3.09%). The reason, we argue, is that the class-agnostic attention is only able to distinguish the dynamic actions from the background frames (e.g., solving the task in THUMOS 14 as shown in Tab. 1). It fails if the mixed actions are all semantically meaningful video frames.

**Contributions of Multimodal Features.** We analyze the contributions of multimodal features by building our models on different subsets of features. We first build our models using single modalities, then present our model considering all types of features.

Tab. 3 shows the results. Among the single modality models, the RGB model provides the best performance on Action Detection (mAP 3.02%). It achieves both high verb detection (mAP 5.75%) and high noun detection (mAP 5.48%) performance. The flow model (action mAP 1.54%) is worse than the RGB model but is better than the Audio model. We can see that the flow feature provides more information for the dynamic actions (verb mAP 4.30%), while it is not that good at localizing static objects temporally (noun mAP 2.61%). The audio model (action mAP 0.20%) is the worst among the three single modal models, but it still provides useful information, especially in verb detection (mAP 1.03%).

Our final model takes advantage of all features and achieves the best performance in terms of action detection mAP (3.19%). Compared to the RGB model, it utilizes the flow and audio information to better detect the dynamic actions (verb mAP 6.43% v.s. 5.75%). Furthermore, compared to the flow and audio models, it combines the appearance feature (RGB) to better recognize objects in the temporal domain (noun mAP 5.22% v.s. 2.61%, 0.50%). In sum, we conclude that our modeling of the videos' multimodal nature helps improve the weakly supervised action detection task.

We show in Tab. 4 and Fig. 4 the verb and noun classes best detected by the three modalities. For the verb detection (Fig. 4 (left)), action "wash" can be easily detected by all three modalities, while "fold" only makes a slight sound, so it is hard to recognize by audio. In comparison, "season"

| Verb Detection | | | | | | Noun Detection | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RGB | | Flow | | Audio | | RGB | | Flow | | Audio | |
| Name | AP(%) | Name | AP(%) | Name | AP(%) | Name | AP(%) | Name | AP(%) | Name | AP(%) |
| wash | 39.59 | wash | 37.14 | wash | 17.47 | corn | 33.21 | yoghurt | 28.50 | microwave | 18.33 |
| filter | 31.33 | hang | 29.69 | season | 11.81 | raisin | 33.17 | tray | 21.81 | salt | 11.79 |
| rip | 30.55 | fold | 23.32 | measure | 8.12 | yoghurt | 33.17 | lid | 21.14 | oatmeal | 5.63 |
| season | 30.11 | dry | 19.88 | unscrew | 6.65 | olive | 29.34 | cloth | 20.72 | carrot | 5.27 |
| fold | 25.51 | throw | 17.75 | squeeze | 5.11 | lid | 28.16 | oven | 18.67 | cupboard | 4.31 |

Table 4. Top-5 classes detected by the RGB, Flow, and Audio features.

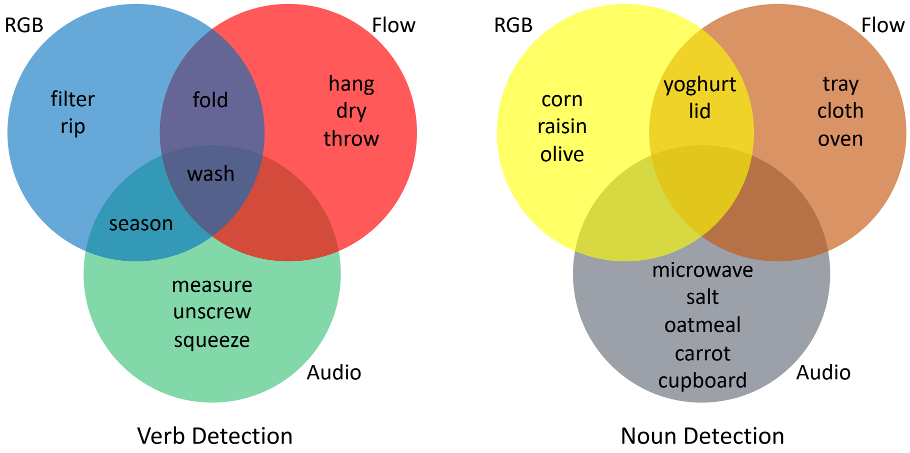

Figure 4. Venn diagrams - The easily detected top-5 classes by different modalities.

sounds loud, but the dynamic action is nuanced; thus, the audio can detect it but not the motion flow. The noun detection results are also interesting (Fig. 4 (right)). We found the "tray" and "cloth" to be more dynamic, and we notice that "microwave" makes a sound. So, we conclude that different modalities help localize different objects and actions temporally.

### 4.3. Qualitative Examples

We provide a qualitative example visualizing the results of our model. Fig. 5 shows it confidently and correctly localizes the actions "wash pan", "wash spatula", and "wash plate". For actions such as "pour liquid:washing" and "wash sponge", our model's estimations of the starting and ending time are not precise, thus causing the IoU with the ground truth to be smaller than 0.5. We can hardly find mistakes regarding classification issues in the top-20. Hence we conclude that localization and refining the action boundaries are still challenging for weakly supervised action detection and should gather more attention.

## 5. Conclusion

We explored audio narration as a form of supervision in this paper. We developed a model to learn from the narration supervision and utilize multimodal features, including RGB, motion flow, and ambient sound. In our design, the model learns to attend to the frames related to the narration label while suppressing the irrelevant frames from being used. In the experiments, we show that the proposed method outperformed alternative designs. Also, we proved that the different modalities contribute to the detections of different actions and objects in the temporal domain.

The insights of our paper are interesting. Throughout the EPIC Kitchen tasks C1-C5, none of them directly uses the cheap audio narration supervision to learn action detectors, while we proved such a task of using narration to be possible. Our experiments have shown that it is plausible to eliminate the expensive stages of refining action boundaries during video detection data annotation. Further, the refined instance-level annotations did not contribute too much to the detector's performance. We expect weakly-, semi-, and self-supervised methods to gather more attention in future video detection tasks.

**Acknowledgement:** This work was supported by a University of Pittsburgh CS50 Computer Science fellowship, and by National Science Foundation award 2046853.

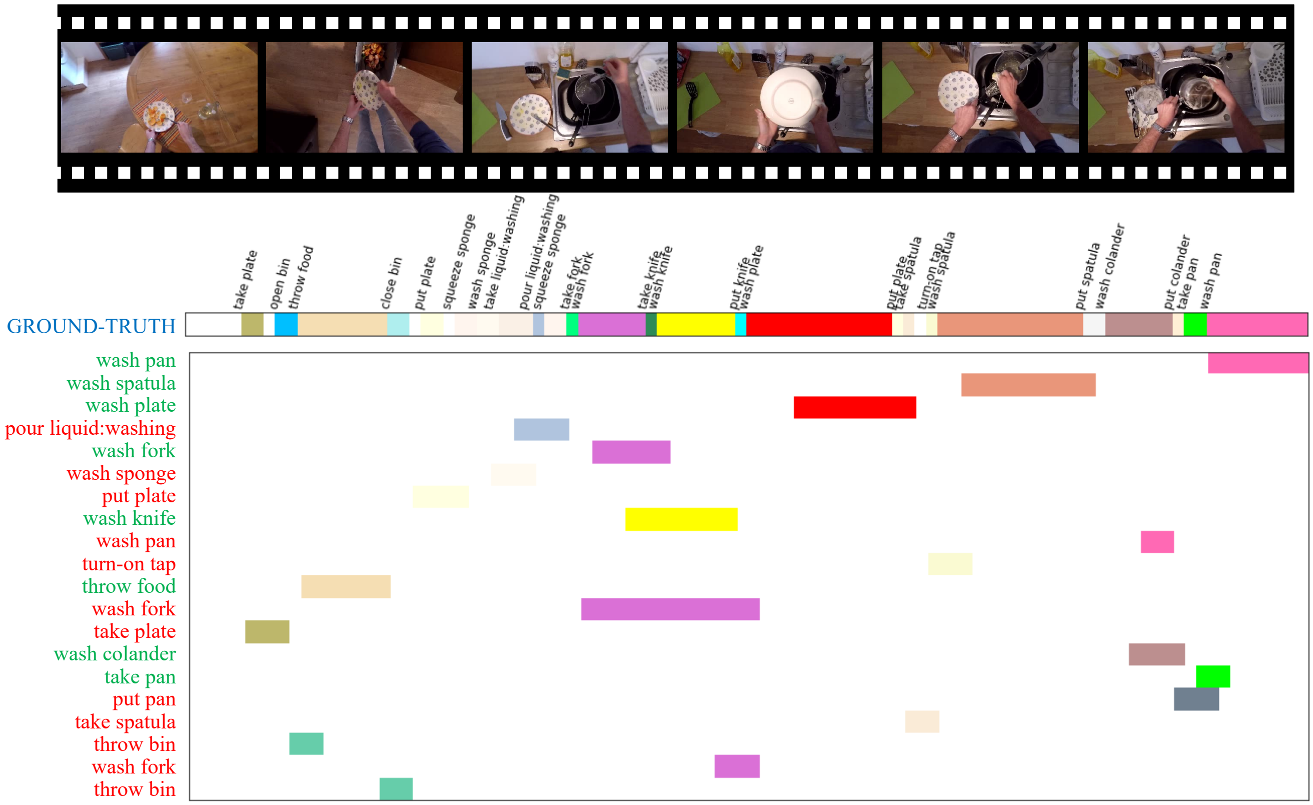

Figure 5. Qualitative example of our model's action detection results. We show the demo of the video, the ground-truth annotations, and our model's top-20 predictions. We show the correct predictions using green and incorrect ones using red. The correctness is determined by IoU@0.5.

## References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 5

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. 2

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, 2016. 2

[6] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[8] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[9] Changjian Chen, Jing Wu, Xiaohan Wang, Shouxing Xiang, Song-Hai Zhang, Qifeng Tang, and Shixia Liu. Towards better caption supervision for object detection. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 2

[10] Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. Amc: Attention guided multi-modal correlation learning for image search. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[11] 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. *arXiv preprint arXiv:2006.13256*, 2020. 5, 6

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 3, 4, 5

[13] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[14] 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 (ICCV)*, October 2019. 1, 3, 5

[15] 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 (ICCV)*, October 2019. 4

[16] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 4, 5

[18] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 3

[19] David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[20] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2

[21] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23, 2017. 2

[22] Mihir Jain, Jan van Gemert, Herve Jegou, Patrick Bouthemy, and Cees G.M. Snoek. Action localization with tubelets from motion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3

[23] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, et al. T-cnn: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2896–2907, 2017. 3

[24] Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Object detection from video tubelets with convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[25] Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, October 2016. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations (ICLR)*, 2015. 5

[27] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 3

[28] Krishna Kumar Singh, Fanyi Xiao, and Yong Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[29] Suha Kwak, Minsu Cho, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Unsupervised object discovery and tracking in video collections. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[30] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International Conference on Learning Representations*, 2020. 3

[31] Zhe Li, Yazan Abu Farha, and Jurgen Gall. Temporal action segmentation from timestamp supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8365–8374, June 2021. 3

[32] 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)*, October 2019. 1, 5

[33] 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)*, September 2018. 1

[34] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 3

[35] Ishan Misra, Abhinav Shrivastava, and Martial Hebert. Watch and learn: Semi-supervised learning for object detectors from video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 3

[36] Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. Weakly supervised video moment retrieval from text queries. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[37] Davide Moltisanti, Sanja Fidler, and Dima Damen. Action recognition from single timestamp supervision in untrimmed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[38] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, 2011. 2

[39] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3

[40] Phuc Xuan Nguyen, Deva Ramanan, and Charless C. Fowlkes. Weakly-supervised action localization with background modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 3

[41] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[42] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 3

[43] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[44] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 3

[45] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 3

[46] Khurram Soomro, Haroon Idrees, and Mubarak Shah. Action localization in videos through context walk. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 1, 3

[47] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Loddon Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[48] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[49] Guiyu Tian, Shuai Wang, Jie Feng, Li Zhou, and Yadong Mu. Cap2seg: Inferring semantic and spatial context from captions for zero-shot image segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4125–4134, 2020. 2

[50] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 3

[51] Daniel R Vilar and Claudio A Perez. Extracting structured supervision from captions for weakly supervised semantic segmentation. *IEEE Access*, 9:65702–65720, 2021. 2

[52] Fang Wan, Pengxu Wei, Jianbin Jiao, Zhenjun Han, and Qixiang Ye. Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[53] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 3

[54] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[55] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[56] Keren Ye, Kyle Buettner, and Adriana Kovashka. Story understanding in video advertisements. In *British Machine Vision Conference (BMVC)*, page 57. BMVA Press, 2018. 3

[57] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8289–8299, June 2021. 2

[58] Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[59] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 3

[60] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14393–14402, June 2021. 2

[61] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 3

[62] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1823–1834, October 2021. 2