CROSS-ATTENTIONAL AUDIO-VISUAL FUSION FOR WEAKLY-SUPERVISED ACTION LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporally localizing actions in videos is one of the key components for video understanding. Learning from weakly-labelled data is seen a potential solution towards avoiding expensive frame-level annotations. Different to others works, which only depend on the visual-modality, we propose to learn richer audio-visual representations for weakly-supervised action localization. First, we propose a multi-stage cross-attention mechanism to collaboratively fuse audio and visual features, which preserves the intra-modal characteristics. Second, to model both foreground and background frames, we construct an open-max classifier, which treats the background class as an open-set. Third, for precise action localization, we design consistency losses to enforce temporal continuity for the action-class prediction, and also help with foreground-prediction reliability. Extensive experiments on two publicly available video-datasets (AVE and ActivityNet1.2) show that the proposed method effectively fuses audio and visual modalities, and achieves state-of-the-art results for weakly-supervised action localization.

1 INTRODUCTION

The goal of this paper is to temporally localize actions and events of interest in videos with weak-supervision. In the weakly-supervised setting, only video-level labels are available during the training phase to avoid expensive and time-consuming frame-level annotation. This task is of great importance for video analytics and understanding. Several weakly-supervised methods have been developed for it (Nguyen et al., 2018; Paul et al., 2018; Narayan et al., 2019; Shi et al., 2020; Jain et al., 2020) and considerable progress has been made. However, only visual information is exploited for this task and audio modality has been mostly overlooked. Both, audio and visual data often depict actions from different viewpoints (Guo et al., 2019). Therefore, we propose to explore the joint audio-visual representation to improve the temporal action localization in videos.

A few existing works (Tian et al., 2018; Lin et al., 2019; Xuan et al., 2020) have attempted to fuse audio and visual modalities to localize *audio-visual events*. These methods have shown promising results, however, these audio-visual events are essentially actions that have strong audio cues, such as playing guitar, and dog barking. Whereas, we aim to localize wider range of actions related to sports, exercises, eating etc. Such actions can also have weak audio aspect and/or can be devoid of informative audio (e.g. with unrelated background music). Therefore, it is a key challenge to fuse audio and visual data in a way that leverages the mutually complementary nature while maintaining the modality-specific information.

In order to address this challenge, we propose a novel multi-stage cross attention mechanism. It progressively learns features from each modality over multiple stages. The inter-modal interaction is allowed at each stage only through cross attention, and only at the last stage are the visually-aware audio features and audio-aware visual features are concatenated. Thus, an audio-visual feature representation is obtained for each snippet in videos.

Separating background from actions/events is a common problem in temporal localization. To this end, we also propose: (a) foreground reliability estimation and classification via open-max classifier and (b) temporal continuity losses. First, for each video snippet, an open-max classifier predicts scores for action and background classes, which is composed of two parallel branches for action classification and foreground reliability estimation. Second, for precise action localization with weak supervision, we design temporal consistency losses to enforce temporal continuity of action-class prediction and foreground reliability.

We demonstrate the effectiveness of the proposed method for weakly-supervised localization of both audio-visual events and actions. Extensive experiments are conducted on two video datasets for localizing audio-visual events (AVE[1]) and actions (ActivityNet1.2[2]). To the best of our knowledge, it is the first attempt to exploit audio-visual fusion for temporal localization of unconstrained actions in long videos.

2 RELATED WORK

Weakly-supervised action localization: In (Wang et al., 2017) and (Nguyen et al., 2018), multiple instance learning scheme (Dietterich et al., 1997) along with attention mechanism was used to localize the actions in videos. Paul et al. (2018) introduced a co-activity similarity loss that looks for similar temporal regions in a pair of videos containing a common action class. Narayan et al. (2019) proposed center loss for the discriminability of action categories at the global-level and counting loss for separability of instances at the local-level. To alleviate the confusion due to background (non-action) segments, Nguyen et al. (2019) developed the top-down class-guided attention to model background, and (Yu et al., 2019) exploited temporal relations among video segments. Jain et al. (2020) segmented a video into interpretable fragments, called ActionBytes, and used it effectively for action proposals. To distinguish action and context (near-action) snippets, (Shi et al., 2020) designed the class-agnostic frame-wise probability conditioned on the attention using conditional variational auto-encoder. Luo et al. (2020) proposed an expectation-maximization multi-instance learning framework where the key instance is modeled as a hidden variable

Audio-visual event localization: The task of audio-visual event localization, as defined in the literature, is to classify each time-step into one of the event classes or background. This is different from action localization, where the goal is to determine the start and the end of each instance of the given action class. In (Tian et al., 2018), a network with audio-guided attention was proposed, which showed prototypical results for audio-visual event localization, and cross-modality synchronized event localization. To utilize both global and local cues in event localization, Lin et al. (2019) conducted audio-visual fusion in both of video-level and snippet-level using multiple LSTMs. Assuming single event videos, (Wu et al., 2019) determined the event class using a video-level feature in one modality, and detected the event-related snippet by matching the video-level feature with snippet-level features from the other modality. In order to address the temporal inconsistency between audio and visual modalities, (Xuan et al., 2020) devised the modality sentinel, which filters out the event-unrelated modalities. Encouraging results have been reported, however, the localization capability of these methods has been shown only for the short fixed-length videos with distinct audio cues. Differently, we aim to also localize actions in long, untrimmed and unconstrained videos.

Deep multi-modal representation learning: Multi-modal representation learning methods aim to obtain powerful representation ability from multiple modalities (Guo et al., 2019). With the advancement of deep-learning, many deep multi-modal representation learning approaches have been developed. Several methods fused features from different modalities in a joint subspace by out-product (Zadeh et al., 2017), bilinear pooling (Fukui et al., 2016), and statistical regularization (Aytar et al., 2017). In other types of methods, features of each modality were disjointly learned under cross-modality constraints such as cross-modal ranking (Frome et al., 2013; Lazaridou et al., 2015; Kiros et al., 2014) or feature distance (Pan et al., 2016; Xu et al., 2015; Liong et al., 2016). Additionally, the encoder-decoder framework was also exploited. Mor et al. (2018) proposed a multi-domain wavenet autoencoder where the encoder is shared across different modalities, and the decoder is designed for each modality to produce musical translations. Contrarily, in (Huang et al., 2018), a network uses multiple encoders to decompose the source modality into modality-specific and modality-invariant subspaces for image-to-image translation.

[1]https://github.com/YapengTian/AVE-ECCV18
[2]http://activity-net.org/download.html

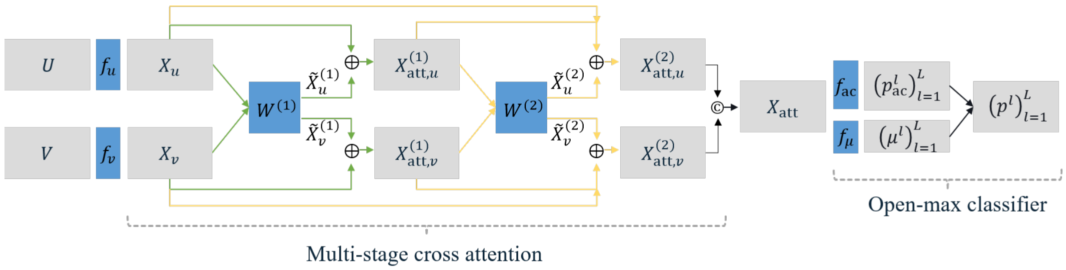


Figure 1: **The proposed architecture** has two parts: modality fusion and open-max classification. (a) *Fusion by Multi-stage cross attention*: The input audio (U) and visual (V) features are embedded by the two fully-connected layers f_u and f_v and passed through the multiple stages of cross-attention. At t-th stage, the attended audio-visual embeddings, $\tilde{X}^{(t)}_{att,u}$ and $\tilde{X}^{(t)}_{att,v}$, are calculated using the results from the previous stages. Here, ⓒ and $\oplus$ denote concatenation and summation operations. The dense skip connections of two stages are depicted as green and yellow arrows. At the last stage, two attended features are concatenated. (b) *Open-max classifier* takes the concatenated audio-visual features as input and generates classification scores for action classes and background.

3 METHODOLOGY

In this section, we introduce the proposed framework for weakly-supervised action and event localization. Fig. 1 illustrates the complete framework. We first present the multi-stage cross attention mechanism to generate audio-visual features in Sec. 3.1. Then, we explain open-max classification to robustly distinguish actions[3] from unknown background in 3.2. Finally, in 3.3, we describe the training loss including two consistency losses designed to enforce temporal continuity of actions and background.

Problem statement: We suppose that a set of videos only with the corresponding video-level labels are given for training. For each video, we uniformly sample L non-overlapping snippets, and then extract the audio features $U = (u^l)_{l=1}^L \in \mathbb{R}^{d_u \times L}$ with a pre-trained network, where u^l is the d_u dimensional audio feature representation of the snippet l. Similarly, the snippet-wise visual features $V = (v^l)_{l=1}^L \in \mathbb{R}^{d_v \times L}$ are also extracted. The video-level label is represented as $c \in \{0, 1, \dots, C\}$, where C is the number of action classes and 0 denotes the background class. Starting from the audio and visual features, our approach learns to categorize each snippet into $C + 1$ classes and hence localizes actions in weakly-supervised manner.

3.1 MULTI-STAGE CROSS ATTENTION MECHANISM

While multiple modalities can provide more information than a single one, the modality-specific information may be reduced while fusing different modalities. To reliably fuse the two modalities, we develop a *multi-stage cross attention* mechanism where features are separately learned for each modality under constraints from the other modality. In this way, the learned features for each modality encodes the inter-modal information, while preserving the exclusive and meaningful intra-modal characteristics.

As illustrated in Fig. 1, we first encode the input features U and V to $X_u = (x_u^l)_{l=1}^L$ and $X_v = (x_v^l)_{l=1}^L$ via modality-specific fully-connected (FC) layers f_u and f_v, where x_u^l and x_v^l are in $\mathbb{R}^{d_x}$. After that, we compute the cross-correlation of X_u and X_v to measure inter-modal relevance. To reduce the gap of the heterogeneity between the two modalities, we use a learnable weight matrix $W \in \mathbb{R}^{d_x \times d_x}$ and compute the cross-correlation as

$$\Lambda = X_u^T W X_v \tag{1}$$

where $\Lambda \in \mathbb{R}^{L \times L}$. Note that x_u^l and x_v^l are l_2-normalized before computing the cross-correlation.

In the cross-correlation matrix, a high correlation coefficient means that the pair of the corresponding audio and visual snippet features are highly relevant. Accordingly, the l^{th} column of Λ corresponds

[3]For brevity we refer both action and event as action.

to the relevance of x_v^l to L audio snippet features. Based on this, we generate cross attention weights A_u and A_v by column-wise soft-max of Λ and Λ^T, respectively. Then, for each modality, the attention weights are used to reweight the snippet features to make them more discriminative in consideration of relationship with the other modality. Formally, the attention-weighted features $\tilde{X}_u$ and $\tilde{X}_v$ are obtained by

$$\tilde{X}_u = X_u A_u \quad \text{and} \quad \tilde{X}_v = X_v A_v. \tag{2}$$

Note that each modality guides the other one through attention weights. This is to ensure the meaningful intra-modal information is well-preserved while applying the cross attention.

To extensively delve into cross-modal information, we repeatedly apply the cross attention in multiple times. However, during the multi-stage cross attention, the original modality-specific characteristics may be over-suppressed. To prevent this, we adopt the dense skip-connection (Huang et al., 2017). More specifically, at stage t, we obtain the attended audio features by

$$X_{\text{att},u}^{(t)} = \tanh\left(\sum_{i=0}^{t-1} X_{\text{att},u}^{(i)} + \tilde{X}_u^{(t)} \right) \tag{3}$$

where $X_{\text{att},u}^{(0)}$ is X_u, and $\tanh(\cdot)$ denotes the hyperbolic tangent activation function. Similar to $X_{\text{att},u}^{(t)}$, the attended visual features $X_{\text{att},v}^{(t)}$ are generated for the visual modality.

At the last stage t_e, we concatenate the attended audio and visual features to yield audio-visual features,

$$X_{\text{att}} = [\, X_{\text{att},u}^{(t_\text{e})};\, X_{\text{att},v}^{(t_\text{e})} \,] \tag{4}$$

where t_e is empirically set to 2 which will be discussed in the ablation studies in Section 4.3.

3.2 Open-max classification

Video segments can be dichotomized into foreground actions and background. For precise action localization, distinguishing the background from the actions is important as well as categorizing the action classes. However, unlike action classes, the background class comprises of extremely diverse types of non-actions. Therefore, it is not possible to train for the wide range of background classes that the model may confront at the test time.

To resolve this problem, we address the background as an open set (Dietterich, 2017; Bendale & Boult, 2016). As illustrated in Fig. 1, we construct an open-max classifier on top of the multi-stage cross-attentional feature fusion. Specifically, the open-max classifier consists of two parallel FC layers for action classification and foreground reliability estimation. The attended audio-visual feature x_{att}^l, where $l = 1, \ldots, L$, are snippet-wisely fed into the open-max classifier. The first FC layer outputs a snippet-wise activation vector $h^l = [h^l(1), \ldots, h^l(C)]$ for C action classes, which is converted to probability scores, p_{ac}^l by soft-max function.

Simultaneously, the second FC layer is applied on x_{att}^l, followed by a sigmoid function to estimate its *foreground reliability*, $\mu^l \in [0, 1]$. The foreground reliability, μ^l, is the probability of snippet l belonging to any action class. The low reliability indicates that no action occurs in the snippet. Therefore, we compute the probability for the background class as the complement of μ^l, by $p_{\text{bg}}^l = 1 - \mu^l$.

Lastly, the open-max classifier outputs the probability distribution p^l over $C + 1$ classes including the background and C actions as

$$p^l = [p_{\text{bg}}^l;\, \mu^l p_{\text{ac}}^l]. \tag{5}$$

3.3 Training Loss

Next, we describe the loss functions to train our model. The actions or foreground do not abruptly change over time. To impose this constraint, we devise two types of *temporal continuity* losses.

Foreground continuity loss: Foreground continuity implies two important properties for neighboring snippets: (a) similar foreground reliability in a class-agnostic way, and (b) consistent open-max probabilities for a target foreground class.

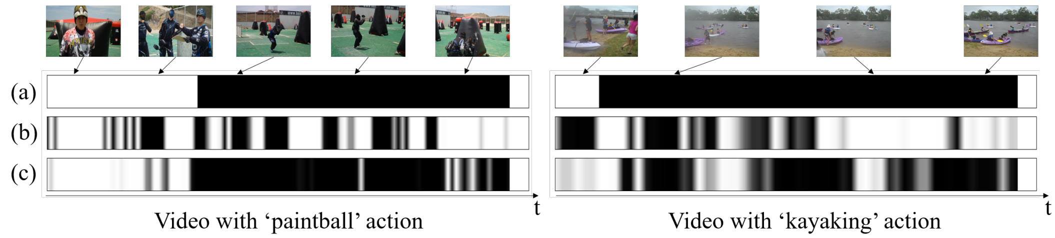


Figure 2: Visualization of class activation sequences for the target actions in two example videos: (a) shows the ground-truth segments, and (b) and (c) show the class activation sequences of the proposed model trained respectively without and with $\mathcal{L}_{\text{cont}}$ and $\mathcal{L}_{\text{pseu}}$. The activation is depicted in gray-scale, where higher intensity indicates more strong activation.

The first of the two constraints is imposed via class-agnostic foreground continuity:

$$\mu_{\text{ag}}^{l} = \frac{1}{B+1} \sum_{i=-B/2}^{B/2} G(i)\, \mu^{l-i} \tag{6}$$

where $G(i)$ is a Gaussian window of width $B+1$ to apply temporal smoothing around μ^{l-i}. For the second constraint, temporal Gaussian smoothing is applied over open-max probability of video-level ground-truth action class, $\hat{c}$, to obtain class-specific foreground continuity:

$$\mu_{\text{sp}}^{l} = \frac{1}{B+1} \sum_{i=-B/2}^{B/2} G(i)\, p^{l-i}(\hat{c}) \tag{7}$$

Finally, the foreground continuity loss is defined as:

$$\mathcal{L}_{\text{cont}} = \frac{1}{L} \sum_{l=1}^{L} |\mu^{l} - \mu_{\text{ag}}^{l}| + |\mu^{l} - \mu_{\text{sp}}^{l}|. \tag{8}$$

The foreground continuity loss imposes temporal continuity of foreground, and hence also helps in separating background from action classes.

Pseudo localization loss: Here, we consider the action or background class continuity, which implies that the open-max probabilities, p^l, agrees with the classification of neighbouring snippets. This can be used to obtain the pseudo label for snippet l. We first average the open-max prediction of N neighbor snippets and itself, $q^l = \frac{1}{N+1} \sum_{i=l-N/2}^{l+N/2} p^i$. We set $\hat{c}^l = \arg\max_c (q^l(c))$ as the pseudo label, but only retain it when the largest class probability of q^l is higher than a predefined threshold τ. Accordingly, the pseudo localization loss is formulated by

$$\mathcal{L}_{\text{pseu}} = \frac{1}{L} \sum_{l=1}^{L} \mathbb{1}(\max(q^l) \geq \tau)(-\log p^l(\hat{c}^l)) \tag{9}$$

Total loss: Additionally, we employ the multiple instance learning (MIL) and co-activity similarity (CAS) losses (Paul et al., 2018). The final loss $\mathcal{L}$ is defined by

$$\mathcal{L} = \mathcal{L}_{\text{mil}} + \alpha \mathcal{L}_{\text{cas}} + \beta \mathcal{L}_{\text{cont}} + \gamma \mathcal{L}_{\text{pseu}} \tag{10}$$

where $\mathcal{L}_{\text{mil}}$ and $\mathcal{L}_{\text{cas}}$ denote MIL and CAS losses, respectively. For details see Appendix C.

Figs. 2 (b) and (c) compare the class activation sequences along the temporal axis for the target classes between the models trained without and with the two consistency losses, respectively. We see that class activations are more continuous in the model with the consistency losses.

4 Experiments

In this section, we provide experimental analysis and comparative evaluation to show the effectiveness of the proposed method.

Table 1: **Ablation for multi-stage cross-attention.** The results for different stages of cross attention are reported for the AVE and ActivityNet1.2 datasets. The comparison with the uni-modal approach shows the impact of leveraging multi-modality and cross-attention.

			Uni-modal		Multi-modal			
			Audio	Visual	0-stage	1-stage	2-stage	3-stage
AVE	Accuracy (%)		32.1	45.2	65.0	75.0	77.1	75.6
ActivityNet1.2	mAP@IoU (%)	0.5	12.3	38.3	37.6	42.1	44.8	39.5
		0.6	10.9	32.9	32.4	35.3	37.8	33.8
		0.7	9.7	25.4	26.7	29.5	30.8	27.9
		0.8	7.6	19.2	19.4	20.8	22.5	20.9
		Avg.	7.8	22.1	22.0	24.1	26.0	23.3

4.1 Datasets and evaluation method

Datasets: We evaluate our approach on Audio-Visual Event (AVE) and ActivityNet1.2 datasets.

AVE dataset is constructed for audio-visual event localization, which contains 3,339 training and 804 testing videos, each lasting 10 seconds with event annotation per second. There are 28 audio-visual event categories covering a wide range of domains, such as animal and human actions, vehicle sounds, and music performance. Each event category has both audio and visual aspects, e.g. church bell, baby crying, man speaking etc.

ActivityNet1.2 is a temporal action localization dataset with 4,819 train and 2,383 validation videos, which in the literature is used for evaluation. It has 100 classes, with on an average 1.5 action instances per video. The average length of the videos in this dataset is 115 seconds.

Evaluation metric: We follow the standard evaluation protocol of each dataset. For the AVE dataset, we report snippet-wise event prediction accuracy. For the ActivityNet1.2 dataset, we generate the action segments (start and end time) from snippet-wise prediction (details are described in the following section), and then measure mean average precision (mAP) at different intersection over union (IoU) thresholds.

4.2 Feature extraction and implementation Details

Feature extraction We use the I3D network (Carreira & Zisserman, 2017) and ResNet152 architecture (He et al., 2016) to extract the visual features for ActivityNet1.2 and AVE, respectively. The I3D network is pre-trained on Kinetics-400 (Kay et al., 2017), and the features consist of two components: RGB and optical flow. The ResNet 152 is pre-trained on the ImageNet (Russakovsky et al., 2015), and the features are extracted from the last global pooling layer. To extract the audio features, we use the VGG-like network (Hershey et al., 2017), which is pre-trained on AudioSet (Gemmeke et al., 2017), for both AVE and ActivityNet1.2 datasets.

Implementation Details We set d_x to 1,024, and the LeakyRelu and hyperbolic tangent functions are respectively used for the activation of modality-specific layers and cross attention modules. In training, the parameters are initialized with Xavier method (Glorot & Bengio, 2010) and updated by Adam optimizer (Kingma & Ba, 2015) with the learning rate of 10^{-4} and the batch size of 30. Also, the dropout with a ratio of 0.7 is applied for the final attended audio-visual features. In the loss, the hyper parameters are set as $B = 4$, $\alpha = 0.8$, $\beta = 0.8$ and $\gamma = 1$.

Localization at test time For event localization at test time, i.e. snippet classification, each snippet l is classified into one of event classes (including background) by $\arg\max_c\, p^l(c)$, where p^l is the open-max probability of snippet l. For action localization, we follow the two-stage thresholding scheme of (Paul et al., 2018). The first threshold is applied to filter out the classes that have video-level scores less than the average over all the classes. The second threshold is applied along the temporal axis to obtain the start and the end of each action instance.

Table 2: **Ablations for the consistency losses and open-max classifier.** *Consistency losses:* The lower part of the table shows the impact of each of the two consistency losses, when used with open-max classifier. *Open-max vs soft-max:* The results for soft-max is also shown, which demonstrates the advantage of foreground/background modelling by open-max classification on both the datasets.

	Method	$\mathcal{L}_{\text{cont}}$	$\mathcal{L}_{\text{pseu}}$	AVE	ActivityNet1.2 [mAP@IoU (%)]				
				Accuracy (%)	0.5	0.6	0.7	0.8	Avg.
Soft-Max	S-I		✓	68.5	39.4	35.7	30.7	19.8	23.8
Open-Max	O-I	✓		64.9	39.9	33.7	23.8	14.3	20.3
	O-II		✓	75.9	44.1	37.4	31.1	22.4	25.7
	O-III	✓	✓	77.1	44.8	37.8	30.8	22.5	26.0

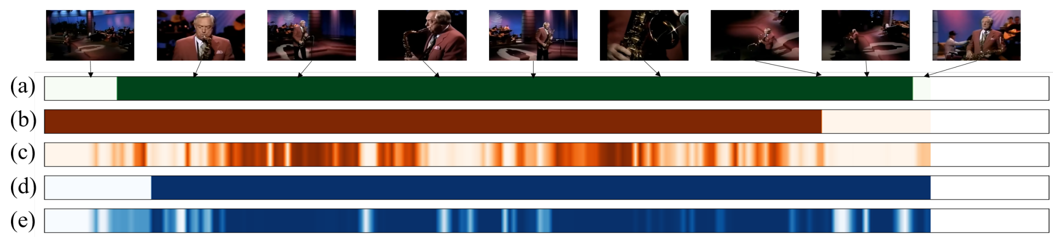


Figure 3: Visualization of the action localization result for an example video from ActivityNet1.2. The ground truth is shown in (a), highlighted in green. The localization and the class activation sequence of the visual-only model are shown in (b) and (c), respectively. Finally, the localization and the class activation sequence for the proposed audio-visual method are shown in (d) and (e).

4.3 Component analysis

Multi-stage cross attention: To evaluate the effectiveness of the multi-stage cross attention in audio-visual fusion, we compare two uni-modal methods (audio or visual) and four multi-modal methods with different stages (0∼3 stages) on the AVE and ActivityNet1.2 datasets in Table 1. The pseudo-label losses and the open-max classifiers are used in all six cases. In the uni-modal methods, the input feature is embedded using an FC layer, and then fed into the open-max classifier. The 0-stage method denotes a naive fusion, where audio and visual features are fused by simple concatenation. with the proposed. Even this naive fusion yields higher performance than the uni-modal methods on the AVE dataset. However, that is not the case with more challenging task of action localization on ActivityNet1.2 dataset. Furthermore, all the later stages improve considerably over 0-stage and the uni-modal cases, for both the datasets. The 2-stage cross attention achieves the best performance for both the dataset. Interestingly, even with the minimal audio cue in ActivityNet1.2 (avg. mAP of audio only is 7.8%), the proposed audio-visual features improved the avg. mAP over visual-only and naive fusion (0-stage) models by 4%.

The Fig. 3 shows the qualitative results of the proposed and visual-only model given an example of ActivityNet1.2 dataset. At the beginning of the video, a performer is shown without any activity. The visual-only model incorrectly predicts the beginning part as a target action while our proposed model correctly predicts it as background. Also, the visual-only model cannot catch the action at the last part of the video since it is visually similar across the frames and has minimal visual activity. Whereas, our model correctly recognizes the last part as an action, owing to the multi-stage cross attention of effective fusion of the two modalities.

Consistency losses: We show the ablation over the two proposed losses, $\mathcal{L}_{\text{cont}}$ and $\mathcal{L}_{\text{pseu}}$, while using Open-Max classifier, in the lower part of the Table 2. The proposed method (O-III) with both the losses performs the best suggesting the importance of both the losses. Further, O-II outperforms O-I by a big margin on both the datasets, implying that the pseudo localization loss is more critical for action localization. This result demonstrates that guiding temporal continuity is essential in the long untrimmed videos as well as the short ones.

Open-max classifier: Lastly, we compare the open-max classifier with soft-max classifier where the last FC layer outputs activations for $C + 1$ classes are normalized by the soft-max function. As the background is considered a closed set in the soft-max approach, the foreground continuity loss is

Table 3: Comparison of the proposed method with the state-of-the-art fully and weakly-supervised methods (separated by '/') on the AVE dataset. Snippet-level accuracy (%) is reported.

Method	Tian et al. (2018)	Lin et al. (2019)	Owens & Efros (2018)	Xuan et al. (2020)	Proposed
Accuracy (%)	74.7 / 73.1	75.4 / 74.2	72.3 / 68.8	77.1 / 75.7	- / **77.1**

Table 4: Comparison of our method with the state-of-the-art action localization methods on the ActivityNet1.2 dataset. The mAPs (%) at different IoU thresholds and average mAP across the IoU thresholds are reported. † indicates audio-visual models. ⋆experiment done using author's code.

Method	Supervision	mAP@IoU (%)										
		0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	Avg.
Zhao et al. (2017)	Full	41.3	38.8	35.9	32.9	30.4	27.0	22.2	18.2	13.2	6.1	26.6
Paul et al. (2018)	Weak	37.0	33.5	30.4	25.7	16.6	12.7	10.0	7.0	4.2	1.5	18.0
Liu et al. (2019b)	Weak	37.1	33.4	29.9	26.7	23.4	20.3	17.2	13.9	9.2	5.0	21.6
Liu et al. (2019a)	Weak	36.8	-	-	-	-	22.0	-	-	-	**5.6**	22.4
Jain et al. (2020)	Weak	39.4	-	-	-	15.4	-	-	-	-	-	-
Shi et al. (2020)	Weak	41.0	37.5	33.5	30.1	26.9	23.5	19.8	15.5	**10.8**	5.3	24.4
Luo et al. (2020)	Weak	37.4	-	-	-	23.1	-	-	-	2.0	-	20.3
Tian et al. (2018)†⋆	Weak	15.4	13.9	12.5	11.2	10.2	9.1	7.6	5.7	1.6	0.3	8.76
Naive fusion (0-stage)†	Weak	39.4	37.0	33.5	30.6	27.7	23.6	20.0	13.6	3.0	0.6	22.9
Ours†	Weak	**44.8**	**42.1**	**37.8**	**34.2**	**30.8**	**26.7**	**22.5**	**15.9**	4.0	1.0	**26.0**

not available. The soft-max is denoted by S-I in Table 2. Both O-II and O-III versions of open-max outperform the S-I method with soft-max. The O-III method improves the accuracy by 8.6% on the AVE dataset and Avg. mAP by 2.2% on the ActivityNet1.2 dataset. This shows the advantsge of modelling background with the open-max classifier.

4.4 Comparison with the state-of-the-art

Audio-visual event localization: In Table 3, we compare the proposed method with the recent fully and weakly-supervised methods on the AVE dataset for audio-visual event localization task. In the weakly-supervised setting, our method performs better than all other existing methods at least by 1.4%. Note that, even though learned in weak-supervision, our approach achieves a comparable accuracy (77.1%) to the state-of-the-art fully-supervised method (Xuan et al., 2020).

Temporal action localization: In Table 4, we apply the proposed method to weakly-supervised action localization in long duration videos of ActivityNet1.2 dataset. The mAP scores at varying IoU thresholds are compared with the current state-of-the-art methods. Our method achieves the highest mAPs for 8 out of 10 IoU thresholds, and outperforms the other methods with the Avg. mAP of 26.0%. We also significantly outperform the audio-visual based method of Tian et al. (2018) and the naive fusion by Avg. mAP of 17.24% and 3.1%, respectively. This demonstrates that the effective fusion of audio and visual modalities is critical for action localization. Furthermore, our approach is even comparable to the fully-supervised method in (Zhao et al., 2017).

5 Conclusion

We present a novel approach for weakly-supervised temporal action localization in videos. In contrast to other methods, we leverage both audio and visual modalities for this task. This is the first attempt at audio-visual localization of unconstrained actions in long videos. To collaboratively fuse audio and visual features, we develop a multi-stage cross attention mechanism that also preserves the characteristics specific to each modality. We propose to use the open-max classifier to model the action foreground and background, in absence of temporal annotations. Our model learns this via two consistency losses that enforce continuity for foreground reliability and open-max probabilities for action classes and the background. In experiments, we demonstrate the importance of each of the proposed components and outperform the state-of-the-art results on both AVE and ActivityNet1.2 datasets.

REFERENCES

Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Cross-modal scene networks. In *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 40, pp. 2303–2314. 2017. 2

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *CVPR*. 2016. 4

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*. 2017. 6

Thomas G. Dietterich. Steps toward robust artificial intelligence. In *AI Mag.*, volume 38, pp. 3–24. 2017. 4

Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. In *Artif. Intell.*, volume 89, pp. 31–71. 1997. 2

Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. DeVise: A deep visual-semantic embedding model. In *NIPS*. 2013. 2

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *arXiv preprint arXiv:1606.01847*. 2016. 2

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 *ICASSP*. 2017. 6

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*. 2010. 6

Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. In *IEEE Access*, volume 7, pp. 63373–63394. 2019. 1, 2

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*. 2016. 6

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *ICASSP*. 2017. 6

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*. 2017. 4

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*. 2018. 2

Mihir Jain, Amir Ghodrati, and Cees G. M. Snoek. ActionBytes: Learning from trimmed videos to localize actions. In *CVPR*. 2020. 1, 2, 8

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. In *arXiv preprint arXiv:1705.06950*. 2017. 6

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*. 2015. 6

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *NIPS*. 2014. 2

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *HLT-NAACL*. 2015. 2

Y. Lin, Y. Li, and Y. F. Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*. 2019. 1, 2, 8

Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Deep coupled metric learning for cross-modal matching. In *IEEE Trans. Multimedia*, volume 19, pp. 1234–1244. 2016. 2

Daochang Liu, Tingting Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*. 2019a. 8

Ziyi Liu, Le Wang, Qilin Zhang, Zhanning Gao, Zhenxing Niu, Nanning Zheng, and Gang Hua. Weakly supervised temporal action localization through contrast based evaluation networks. In *ICCV*. 2019b. 8

Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *ECCV*. 2020. 2, 8

Noam Mor, Lior Wolf, Adam Polyak, and Yaniv Taigman. A universal music translation network. In *arXiv preprint arXiv:1805.07848*. 2018. 2

Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao. 3C-Net: Category count and center loss for weakly-supervised action localization. In *ICCV*. 2019. 1, 2

Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *CVPR*. 2018. 1, 2

Phuc Xuan Nguyen, Deva Ramanan, and Charless C. Fowlkes. Weakly-supervised action localization with background modeling. In *ICCV*. 2019. 2

Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*. 2018. 8

Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*. 2016. 2

Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. In *ECCV*. 2018. 1, 2, 5, 6, 8, 13

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. In *Int. J. Comput. Vis.*, volume 115, pp. 211–252. 2015. 6

Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *CVPR*. 2020. 1, 2, 8

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*. 2018. 1, 2, 8

Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. UntrimmedNets for weakly supervised action recognition and detection. In *CVPR*. 2017. 2

Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*. 2019. 2

Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*. 2015. 2

Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI*. 2020. 1, 2, 8

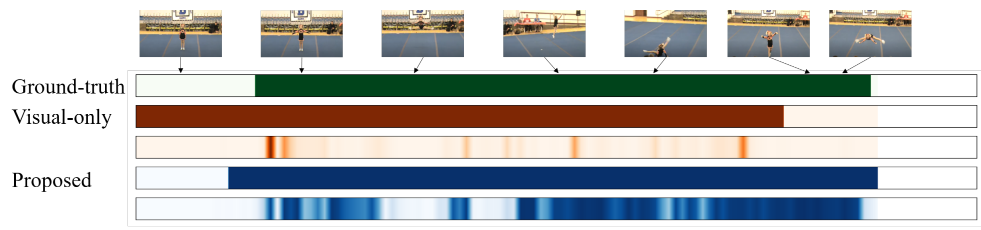
Tan Yu, Zhou Ren, Yuncheng Li, Enxu Yan, Ning Xu, and Junsong Yuan. Temporal structure mining for weakly supervised action detection. In *ICCV*. 2019. 2

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*. 2017. 2

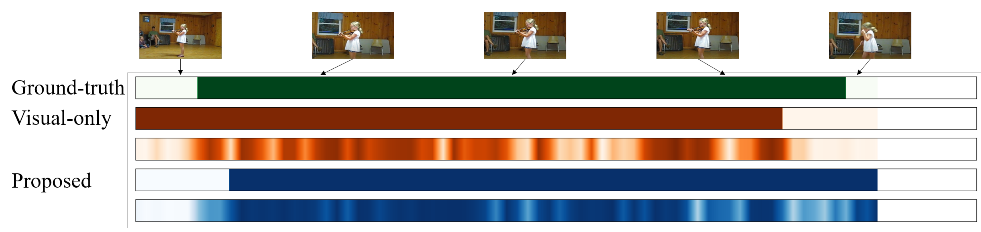
Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*. 2017. 8

A QUALITATIVE EVALUATION

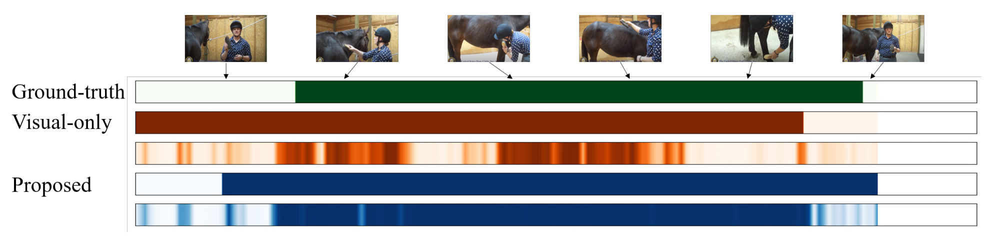
We provide additional qualitative results for action localization on the ActivityNet1.2 dataset. Fig. 4 compares the proposed method with the method trained on visual modality ('Visual-only'). The open-max classifier and total loss function are commonly used for both. In Figs. 4(a) and (b), because the videos are static in visual modality, the background segments in early parts of videos are miss-localized as actions in the visual-only model. Contrarily, proposed method distinguishes the background based on the action-related audio (cheerleading music and violin sound). In Fig. 4(c), the brushing sound is overlapped with the loud human narration lasting throughout videos. Nevertheless, the proposed method effectively extracts the crucial audio cues and fuses them with the visual ones. In Fig. 4(d), even though the early part of the action is visually occluded by large logos, our method exactly localizes the action. Also, for all of the class activation sequences, the activations by the proposed method are more consistently high for actions. This means that our collaboration of audio and visual modalities is more robust in distinguishing foreground from background.



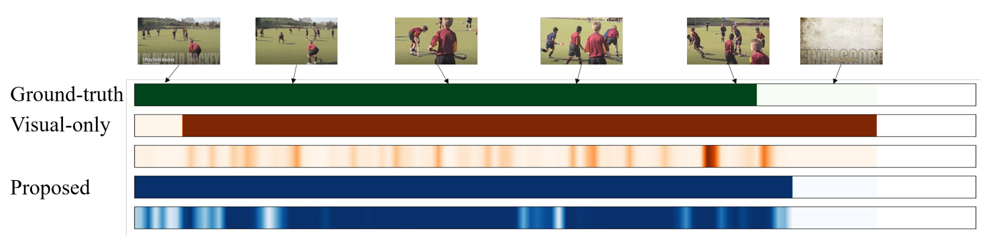
(a) Cheerleading



(b) Playing violin



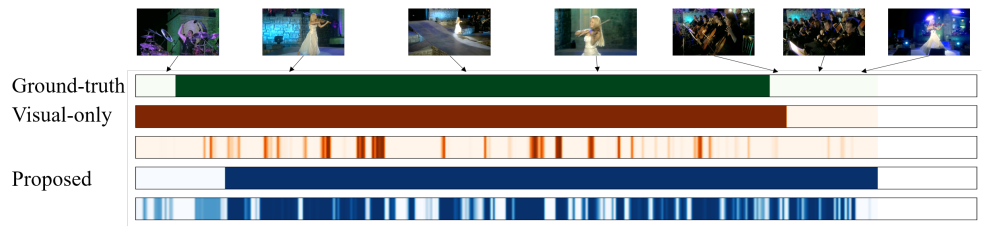
(c) Grooming horse



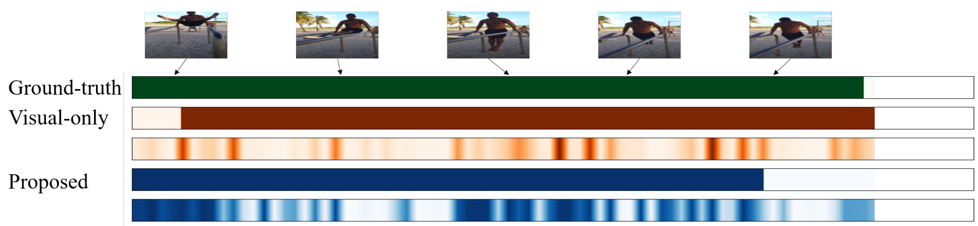
(d) Playing field hockey

Figure 4: Qualitative results for action localization. Ground-truth (green), prediction by the visual-only method (orange), and prediction by the proposed method (blue) are shown. Class activation sequences are visualized below each prediction, darker shade means higher activation.

Fig. 5 illustrates the cases where audio degrades the performance. Fig. 5 (a) shows an example video for action class 'playing violin'. The violin sound of the soloist and the band is intermingled in the video. In the end, the sound of violin continues making our model predict the action but since camera focuses on the band, the ground-truth does not include those frames. Fig. 5 (b) shows an example of action 'using parallel bars'. Here the repeated background music is irrelevant to action, therefore the class activation is bit off in the last part. However, thanks to the visual modality, the prediction is still reasonable.



(a) Due to setting ground-truth confined to visual modality



(b) Due to background music repeated for a long time

Figure 5: Examples where localization performance is degraded by audio.

B Analysis for consistency losses and open-max classification

In Table 5, we conduct more extensive analyses for the consistency losses and open-max classifier. Specifically, we replace the open-max classification approach with soft-max one. Then, for both classifiers with the 2-stage cross attention, we ablate the foreground continuity or pseudo localization losses where CAS and MIL losses are commonly used. First, the performance gap between S-0 and O-0, where only CAS and MIL losses are used, shows the difficulty of learning two parallel branches in weakly-supervised manner. However, when adding the pseudo localization loss, (S-I and O-II), the open-max classification approach is further improved than the soft-max. Hence, pseudo labels that reduce the fallacious actions classification are more effective on the open-set background modeling than the closed-set modeling.

Next, O-I and O-II shows higher performance than O-0. Similarly, S-I is superior to S-0. This indicates that erroneous detection is suppressed by the correctly detected neighbors when using the consistency losses. Also, in comparing O-I and O-II, the pseudo localization loss gives more performance improvement. This is because pseudo localization loss addresses the consistency in entire actions and background classes. In most cases, O-III, open-max classification with both consistency losses, yields the highest performance. Therefore, all of the proposed open-max classification and consistency losses are effective to temporal action or event localization in videos.

Table 5: Ablation analysis on consistency losses for soft-max (S-0 and S-I) and open-max (O-0, O-I, O-II, and O-III) classifiers on the AVE and the ActivityNet1.2 datasets. O-III is the proposed.

Method	$\mathcal{L}_{\text{cont}}$	$\mathcal{L}_{\text{pseu}}$	AVE	ActivityNet1.2 [mAP@IoU (%)]				
			Accuracy (%)	0.5	0.6	0.7	0.8	Avg.
S-0			62.1	36.4	28.4	22.7	15.8	19.6
S-I		✓	68.5	39.4	35.7	30.7	19.8	23.8
O-0			60.4	35.4	27.5	22.9	12.7	18.7
O-I	✓		64.9	39.9	33.7	23.8	14.3	20.3
O-II		✓	75.9	44.1	37.4	31.1	22.4	25.7
O-III	✓	✓	77.1	44.8	37.8	30.8	22.5	26.0

C Multiple instance loss and Co-activity similarity loss

We apply multiple-instance learning loss for classification. The prediction score corresponding to a class is computed as the average of its top k activations over the temporal dimension. Co-activity similarity loss (CASL) (Paul et al., 2018) is computed over two snippet sequences from a pair of videos, to have higher similarity when the videos have a common class.