Boosting Positive Segments for Weakly-Supervised Audio-Visual Video Parsing

Kranthi Kumar Rachavarapu
Indian Institute of Technology Madras, India
kranthi.rachavarapu@gmail.com

Rajagopalan A. N.
Indian Institute of Technology Madras, India
raju@ee.iitm.ac.in

Abstract

In this paper, we address the problem of weakly supervised Audio-Visual Video Parsing (AVVP), where the goal is to temporally localize events that are audible or visible and simultaneously classify them into known event categories. This is a challenging task, as we only have access to the video-level event labels during training but need to predict event labels at the segment level during evaluation. Existing multiple-instance learning (MIL) based methods use a form of attentive pooling over segment-level predictions. These methods only optimize for a subset of most discriminative segments that satisfy the weak-supervision constraints, which miss identifying positive segments. To address this, we focus on improving the proportion of positive segments detected in a video. To this end, we model the number of positive segments in a video as a latent variable and show that it can be modeled as Poisson binomial distribution over segment-level predictions, which can be computed exactly. Given the absence of fine-grained supervision, we propose an Expectation-Maximization approach to learn the model parameters by maximizing the evidence lower bound (ELBO). We iteratively estimate the minimum positive segments in a video and refine them to capture more positive segments. We conducted extensive experiments on AVVP tasks to evaluate the effectiveness of our proposed approach, and the results clearly demonstrate that it increases the number of positive segments captured compared to existing methods. Additionally, our experiments on Temporal Action Localization (TAL) demonstrate the potential of our method for generalization to similar MIL tasks.

1. Introduction

Event detection and localization in videos is an important task for video understanding. Event localization using visual frames has attracted a lot of attention [49, 50, 36, 35, 22, 18, 4, 41, 28, 33, 23, 13]. These methods rely only on visual cues and ignore a crucial modality - audio, which plays an integral part in human perception. To address this, there has been increased attention on audio-visual event detection [39, 43], where an event could occur in either of the modalities. Furthermore, audio and visual modalities could aid each other in better event localization.

In this paper, we explore the problem of weakly-supervised Audio-Visual Video Parsing (AVVP) task, where the goal is to detect and localize events using only video-level event labels. Such a formulation is attractive as it forgoes the need for expensive and tedious fine-grained labeling. However, this is a challenging problem due to the absence of segment-level labels during training and the requirement to process multi-modal (audio-visual) data in unconstrained videos with varying scene content.

Most previous works [38, 43, 15, 20, 26, 3] use instance-level MIL technique [12] with attentive pooling to model the video-level labels from the segment-level predictions. Such models are then trained to optimize for the video-level labels. These models are then used to identify the positive segments from segment-level predictions. While recent models have shown promising results, they often fail to identify all positive segments of an event due to the MIL-based objective as it implicitly prioritizes the most discriminative segments that satisfy weak supervision constraints. Consequently, the model selects only a subset of the most discriminative incomplete set of positive segments, which are sufficient for correctly classifying the video.

Recent work by Wu et.al [43] focuses on reducing the uncertainty due to the absence of the modality labels by estimating them from a model trained with video-level weak labels. Inspired by this, we focus on improving temporal localization by capturing all of the positive segments. To achieve this, we explicitly model the number of positive segments in a video (z) and optimize this along with the MIL objective. We show that the number of positive segments in a video follows the Poisson binomial distribution over segment-level event probabilities, which can be computed exactly. We can use the weak-supervision constraint that a positive video must contain at least one positive segment ($z \geq 1$) to provide supervision for z and train the model. However, this approach also faces issues similar to that of other MIL-based techniques in accurately localizing events temporally. To overcome the challenge of weak labels, we propose an iterative optimization approach using the Expectation-Maximization (EM) algorithm by modeling the number of positive segments (z) as the latent variable. In E-Step, we employ the trained model to estimate the number of positive segments in a video. In the M-step, we optimize for the video-level labels through classification. Here, we propose the MIL objective using our Poisson binomial formulation with weakly-supervised constraints. We iteratively optimize for the model parameters and estimate the minimum number of positive segments in a video to improve the performance of the task.

We fix our network architecture to that of the HAN [38] and show that our carefully-designed training strategy yields performance gains. We evaluate our approach on the weakly-supervised AVVP task and show that our proposed method achieves state-of-the-art performance. Here, our method achieves improvement in terms of recall and precision, which indicates that it is able to better capture the positive segments. Moreover, while architectural changes may constrain a method to the particular task at hand, innovative training strategies can generalize to multiple related tasks. To validate this, we evaluate our approach on the Temporal Action Localization (TAL) task, which involves a substantially higher number of instances per video compared to AVVP. We achieve state-of-the-art results, which show that our proposed approach can generalize effectively to other related tasks as well. Our contributions are:

- We propose to explicitly model and maximize the number of positive segments in a video to improve localization under weak supervision. We show that the number of positive segments follows the Poisson binomial distribution and can be computed exactly from segment-level probabilities in a fully differentiable manner.
- Given the absence of explicit supervision on the number of positives within a video, we propose an EM-based optimization and iteratively optimize for the proposed Poisson binomial-based MIL loss to boost the total number of positive segments using a model under weak supervision.
- Our experiments on the AVVP task show that our proposed approach consistently performs favorably over the state-of-the-art methods on various metrics. Additionally, our experimentation on TAL demonstrates its potential for generalization to similar tasks.

2. Related Work

Temporal Action Localization (TAL) aims at localizing actions in visual frames within a video. Many approaches [49, 50, 36, 35, 22, 18, 4] have been proposed to solve this task under full supervision. Recently, many weakly-supervised approaches have been proposed to alleviate the cost of frame-level annotation. Attention-based methods [41, 28, 33, 23, 13] focus on selecting keyframes with high action probability by learning attention mechanisms with additional constraints. Others aim to identify key frames by exploiting the complementary nature of RGB and flow modalities [47, 45], or using the most discriminative parts [37, 51, 27] of a video.

Audio-Visual Event Localization. One way to extend the TAL problem is to require the model to reason about the multiple modalities of information in videos (the visual and audio streams) as well. Audio-Visual Event Localization (AVEL) [39] task aims to localize events that are both audible and visible. Several approaches are proposed [44, 52, 21, 31, 32] to solve this problem in weak supervision. These models implicitly assume that audio and visual modalities are always correlated and temporally aligned. These approaches model temporal dependencies of audio and video segments using LSTMs or attention mechanisms to fuse features from multiple modalities.

Audio-Visual Video Parsing task also aims at localizing events in a multimodal setting *i.e*. using both audio and visual streams. In contrast to AVEL, the task of audio-visual video parsing (AVVP) [38] aims at localizing events that are either audible, visible or both, and classifying them into known categories under weak supervision. Tian *et al*. [38] proposed a Multimodal Multiple Instance learning (MMIL) based hybrid attention network to capture temporal, unimodal, and cross-modal context simultaneously. Lamba *et al*. [15] efficiently utilize cross-modal information, along with self-supervised and adversarial training to learn better representations for improved event localization. Wu *et al*. proposed an improvement over [43] by generating reliable modality labels and optimizing for them using MMIL approach. Lin *et al*. method [20] exploits both the common and diverse event semantics across videos to identify audio or visual events by exploring event co-occurrence across modalities. Recently Mo *et al*. [26] explicitly modeled semantic-aware grouping to learn discriminative multimodal subspaces. In contrast to these approaches, we propose to improve the temporal localization of events by modeling the total number of positive segments within a video. Then, an iterative refinement strategy is introduced to boost the total number of positive segments within a video, starting with a model trained on weak-supervision constraints.

3. Weakly-Supervised Event Prediction

We first describe the problem formulation (§3.1) and then describe the standard instance-level MIL approach for solving this problem (§3.2). We describe our proposed Poisson binomial distribution formulation and our training strategy in §3.3, §3.4, §3.5. We then show in §3.6 that our proposed approach is an EM algorithm. Finally, we discuss how our proposed formulation relates to many of the existing MIL techniques in §3.7.

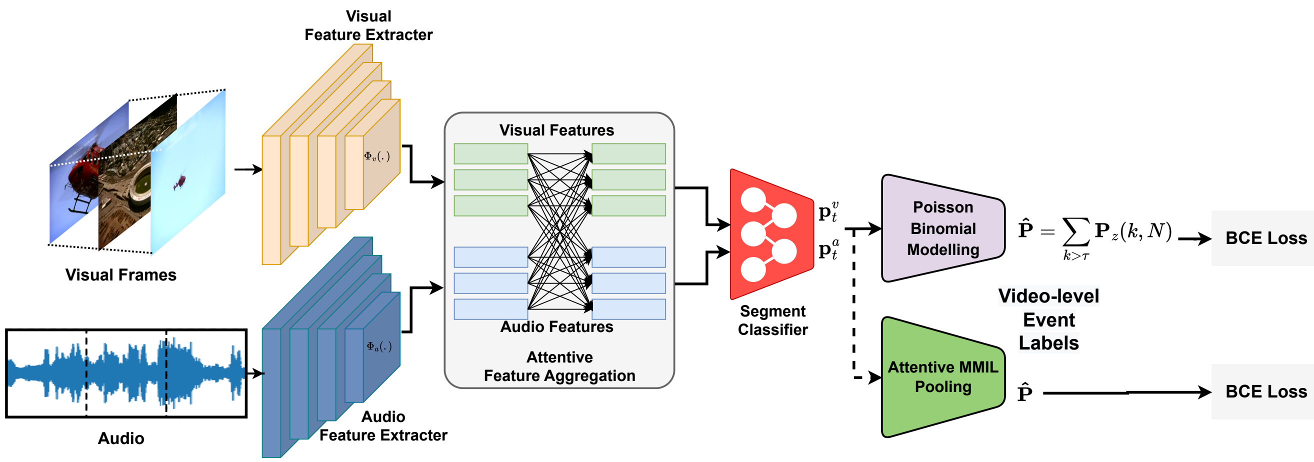

Figure 1: Our proposed audio-visual parsing framework. We employ Hybrid Attention Network (HAN) [38] to predict the segment-level event probabilities ($\hat{\mathbf{p}}_t^a, \hat{\mathbf{p}}_t^v$) from features extracted using self- and cross-attention on audio, visual modalities. Event labels for the entire video are then obtained using an Attentive MMIL pooling and the proposed Poisson Binomial formulation. We propose to model the total positive segments in a video as a Poisson Binomial distribution and compute it exactly from segment-level event probabilities. We then optimize the total positive segments iteratively to improve the event localization of the model.

3.1. Problem Definition

In the audio-visual video parsing (AVVP) problem, our goal is to identify events that are audible or visible in an unconstrained video and simultaneously localize them temporally. Specifically, given a video $\mathbf{X}$ of N non-overlapping temporal segments $\{X_t = (x_t^a, x_t^v)\}_{t=1}^N$ of audio (a), visual (v) streams, our objective is to classify each segment into C possible events (*e.g.* Singing, Vehicle *etc.*). Thus, during evaluation, we need to identify *segment-level* event labels, $\mathbf{y}_t = (y_t^a, y_t^v)$, for each of the audio, visual segments, where $y_t^a, y_t^v \in \{0,1\}^C$ are segment-level audio and visual event labels, respectively. An audio-visual event $y_t^{av} = y_t^a y_t^v$ is defined as an event that is both audible and visible.

Weak Supervision. In weakly-supervised AVVP, for each video $\mathbf{V}$, we only have access to the corresponding video-level event label vector $\mathbf{Y} = [Y_0, Y_1, \ldots, Y_C] \in \{0,1\}^C$, where $Y_c = 1$ if any of the segments in the video contains c-th event, otherwise $Y_c = 0$. These *weak* labels only indicate whether an event occurred in the given video or not. During evaluation, we need to identify *segment*-level labels $\{(y_t^a, y_t^v)\}_{t=1}^N$. Also, note that more than one event can occur in a video, *i.e.* $\sum_c Y_c \geq 1$.

3.2. Overall approach and Architecture

We adopt an instance-level MIL approach [12] where each instance is classified first. These classification scores are then aggregated by MIL pooling. More formally, the video level-label $\hat{\mathbf{P}} \in [0,1]^C$ is generated for a video $X = \{x_t^m\}_{t=1}^N, \forall m \in \{a, v\}$ with T segments as,

$$\hat{\mathbf{P}} = \sigma_{x \in X}\left[g_\phi(f_\theta(x))\right] \quad (1)$$

where, $f_\theta(.)$ is a feature extractor, $g_\phi(.)$ is a segment classifier, $\sigma[.]$ is an MIL-pooling operator, and x is a audio/visual segment in the video X.

In our work, we adopt the Hybrid Attention Network (HAN) architecture of Tian *et al*. [38] to implement Eq. 1. Here, the feature extractor $f_\theta(.)$ consists of Pre-trained audio CNN ($\mathbf{\Phi}_a$), visual CNN ($\mathbf{\Phi}_v$) followed by a two-stream cross-modal transformer ($\mathbf{T}_{av}$). The segment classifier $g_\phi(.)$ is a linear classifier. And, $\sigma[.]$ is an attentive Multi-modal MIL (MMIL) pooling operator. An overview of our method is shown in Figure 1. Please refer to the Supplementary (§S1) for architecture details. During training, the binary cross-entropy (CE) loss between the predicted video-level event probability vector $\hat{\mathbf{P}}$ and the weak video-level label $\mathbf{Y}$ is minimized as,

$$\mathcal{L}_{\text{MIL}}^{\text{Att}} = \mathbf{CE}(\hat{\mathbf{P}}, \mathbf{Y}). \quad (2)$$

During inference, we obtain segment-level predictions for each segment (x_t^m) in a video and threshold them *i.e.* $y_t^m = \mathbb{1}[g_\phi(f_\theta(x_t^m)) \geq 0.5]$ where $\mathbb{1}[.]$ is an indicator function.

Such instance-level MIL approaches enable in finding key segments using the segment classifier. The general weakly-supervised MIL-based formulation for event localization optimizes for a set of most discriminative segments during training and fails to capture all positive segments because of a lack of fine-grained supervision. To address this limitation, we propose to boost the total number of positive segments captured by the model in a video.

3.3. Modelling the number of positive segments using Poisson Binomial Distribution

In AVVP task, the target $y_t^m(c) \in \{0,1\}$ is a binary for each class in AVVP task. Bernoulli distribution is a natural choice for modeling binary classification in ML, where the outcomes are classified as success or failure. This formulation facilitates estimating the success probabilities associated with each label. Therefore, for any given video, we model the segment-level event probabilities as Bernoulli distribution with the success probability of $\hat{\mathbf{p}}_t^m(c) \in [0,1] \;\; \forall t \in [T],\; m \in \{a,v\}, c \in [C]$ for binary classification problem. Thus, the label distribution of all the segments of an event-c are independent and non-identical Bernoulli random variables, as each segment has a different success probability $\hat{\mathbf{p}}_t^m(c)$. To simplify the discussion, we will disregard the event subscript c as we describe our modeling for a class c without any loss of generality. Therefore, for the remainder of the discussion, $\hat{\mathbf{p}}_t^m$ will imply $\hat{\mathbf{p}}_t^m(c)$ and $\mathbf{Y}$ will imply $Y_c \in \{0,1\}$ for the event c under consideration.

Let z be a random variable (RV) denoting the number of *positive* segments with the event in a video *i.e.* $z = \sum_{\forall t,m} \hat{\mathbf{y}}_t^m$, where $\hat{\mathbf{y}}_t^m \sim \text{Bernoulli}(\hat{\mathbf{p}}_t^m)$ indicates whether t-th segment in modality-m has the event or not. The RV z follows the Poisson Binomial distribution, which can be computed exactly from the segment probabilities $\{\hat{\mathbf{p}}_t^m\}_{t=1}^N, m \in \{a,v\}$ as,

$$\hat{\mathbf{P}}_z(k;N) = \tfrac{1}{N+1}\sum_{l=0}^{N} e^{-i\omega lk}\left[\prod_{\forall t,m}(1-\hat{\mathbf{p}}_t^m + \hat{\mathbf{p}}_t^m e^{i\omega l})\right] \quad (3)$$

where, $\omega = \frac{2\pi}{N+1}$. Here, $\hat{\mathbf{P}}_z(k;N)$ is the probability mass function of z and gives the probability of exactly k positive segments and $N-k$ negative segments in a video with N segments. Please refer to Supplementary (§S2) for the derivation. For ease of reference, we define an abbreviated form of Eq. 3 as $\hat{\mathbf{P}}_z(k;N) = \texttt{PoiBin}(\{\hat{\mathbf{p}}_t^m\}_{\forall t,m})$. Therefore, the distribution of RV z, which indicates the number of positive segments, can be modeled explicitly from the segment-level event probabilities. Eq. 3 can be efficiently implemented using 1D-IDFT. Since Eq. 3 is differentiable, it can be effectively utilized in the loss function to enable end-to-end training of the model. The code for implementing Poisson Binomial distribution from segment probabilities is available on our project page[1].

3.4. Poisson Binomial based MIL forumlation

We propose a novel Poisson binomial MIL-pooling operator by modeling the number of positive segments (z) in a video. In contrast to attentive-MIL pooling, our proposed pooling explicitly models the total number of positive segments and thereby aids in improving temporal localization.

In weakly-supervised MIL setup, the only constraint on the segment-level label is that at least one segment (or a portion of segments) in each positive video is positive, and all segments in the negative video are negative. Therefore, the final video-level event probability is computed from the Poisson binomial distribution as,

$$\hat{\mathbf{P}} = \sum_{k \geq \tau} \hat{\mathbf{P}}_z(k;N) \quad (4)$$

where τ is a hyper-parameter indicating the minimum number of positive segments in a video. Eq. 4 describes our choice of the pooling operator $\sigma[.]$ in Eq. 1. Unlike the previous methods that use either pseudo segment-labels [1, 47, 23] or only optimize for weak labels [41, 28, 33, 23, 13, 38, 43, 15, 26] through attentive pooling, we propose to use this z as an intermediate level of supervision for MIL-formulation. We hypothesize that z provides a more informative signal than weak-label while being less noisy than pseudo-segment labels.

Under the weak-supervision constraints, a positive video must contain at least one positive segment. Therefore, we initialize $\tau = 1$ for all videos. We minimize the binary cross-entropy loss between predicted video-level event probability vector $\hat{\mathbf{P}}$ and weak video-level label $\mathbf{Y}$, given by,

$$\mathcal{L}_{\text{MIL}}^{\text{PoiBin}} = \mathbf{CE}(\hat{\mathbf{P}}, \mathbf{Y}) \quad (5)$$

This is equivalent to optimizing the video-level labels when there are at least τ positive segments. As we do not have access to the ground-truth τ during training, optimizing for $k \geq \tau$ includes all possible label assignments. For $\tau = 1$, this approach also faces the same issues as other MIL-based approaches in identifying all the positive segments. Therefore, we use the trained model to get a better estimate of the minimum number of positive segments in a video. We estimate this dynamic threshold for each video clip as,

$$\tau^* = \begin{cases} \arg\max_k \hat{\mathbf{P}}_z(k;N), & \text{if } \mathbf{Y} = 1 \\ 1, & \text{otherwise} \end{cases} \quad (6)$$

Using this new threshold τ^*, we recompute the video-level probability (Eq. 4) and optimize loss (Eq. 5) to retrain the model. The threshold τ^* is computed separately for each event c since the number of positive segments may differ for each event. The complete pseudo-code is given in Algorithm 1. Our proposed approach can be interpreted as an Expectation-Maximization (see §3.6). Consequently, all the convergence guarantees of the EM extend to our proposed iterative approach. We also propose an efficient

[1]https://github.com/KranthiKumarR/poiBin

Algorithm 1: Weakly-Supervised AVVP using the proposed Poisson Binomial based MIL formulation

Initialization: model parameters $\psi = \{\theta, \phi\}$, learning rate β, threshold $\tau_X = 1\ \forall X$,
while *ψ has not converged* **do**
  $\#Train$
  for $(\mathbf{X}, \mathbf{Y}, \tau_X)$ *in train set* **do**
    $\hat{\mathbf{p}}_t^m \leftarrow p_\theta(y|\mathbf{x}_t^m)$;
    $\hat{\mathbf{P}}_z = \texttt{PoiBin}(\{\hat{\mathbf{p}}_t^m\}_{\forall t,m})$;
    $\hat{\mathbf{P}} = \sum_{k \geq \tau_X} \hat{\mathbf{P}}_z(k; N)$;
    $\psi \leftarrow\ \psi - \beta \cdot \nabla_\psi \mathcal{L}(\hat{\mathbf{P}}, \mathbf{Y})$;
  end
  $\#Estimate\ \tau^*$
  for $(\mathbf{X}, \mathbf{Y})$ *in train set* **do**
    $\hat{\mathbf{p}}_t^m \leftarrow p_\theta(y|\mathbf{x}_t^m, \mathbf{Y})$;
    $\hat{\mathbf{P}}_z = \texttt{PoiBin}(\{\hat{\mathbf{p}}_t^m\}_{\forall t,m})$;
    $\tau^* = \arg\max_k \hat{\mathbf{P}}_z(k; N)$;
    Set $\tau_X \leftarrow \tau^*$;
  end
end

way of adapting this approach for multi-class classification in Supplementary (§S3) and employ it in our experiments on TAL task in §5.

3.5. Regularization through Data Augmentation

The Poisson binomial based MIL formulation provides additional flexibility for better estimating the video-level probability, which may not be possible with the attention-based MIL formulations. Consider two training videos $\mathbf{X}^i$ and $\mathbf{X}^j$ with N segments each, and their corresponding video-level weak label sets $\mathbf{Y}^i$ and $\mathbf{Y}^j$. If we create a synthetic video $[\mathbf{X}^i\ \mathbf{X}^j]$ with $2N$ segments, we have three possible cases for any event, and each of them offers a better bound on the threshold τ, as described below.

- $\mathbf{Y^i} \neq \emptyset$ & $\mathbf{Y^i} \cap \mathbf{Y^j} \neq \emptyset$: If both of the videos contain an event, then the event probability for the synthetic video is $\hat{\mathbf{P}} = \sum_{k \geq \tau}^{2N} \hat{\mathbf{P}}_z(k; 2N)$, where $\tau = \tau_i^* + \tau_j^*$.
- $\mathbf{Y^i} \neq \emptyset$ & $\mathbf{Y^i} \cap \mathbf{Y^j} = \emptyset$: Here, only one of the two videos is positive. In such cases, we can get a much better estimate of the video-level event probability as $\hat{\mathbf{P}} = \sum_{k \geq \tau}^{N} \hat{\mathbf{P}}_z(k; 2N)$ with $\tau = \max(\tau_i^*, \tau_j^*)$.
- $\mathbf{Y^i} = \emptyset$ & $\mathbf{Y^i} \cap \mathbf{Y^j} = \emptyset$: Here, none of the videos contain the event. Thus, the video level event probability is $\hat{\mathbf{P}} = \sum_{k \geq \tau}^{2N} \hat{\mathbf{P}}_z(k; 2N)$ with $\tau = 1$

Therefore, during training, we train using the synthetic videos $[\mathbf{X}^i\ \mathbf{X}^j]$ with new thresholds described above to better estimate the video-level event probabilities. Here, we minimize the following loss:

$$\mathcal{L}_{\text{MIL}}^{\text{Aug}} = \mathbf{CE}(\hat{\mathbf{P}}, \mathbf{Y^i} + \mathbf{Y^j}) \tag{7}$$

In summary, we train our model following the Algorithm 1 using the following loss function: $L = \mathcal{L}_{\text{MIL}}^{\text{Att}} + \mathcal{L}_{\text{MIL}}^{\text{PoiBin}} + \mathcal{L}_{\text{MIL}}^{\text{Aug}}$. During inference, we predict the segment-level probabilities $\hat{\mathbf{p}}_t(c)^m = g_\phi(f_\theta(x_t^m)),\ \forall c$, for each segment x_t^m and threshold it to detect all events as $y_t^m(c) = \mathbb{1}\left[\hat{\mathbf{p}}_t^m(c) \geq 0.5\right]$, where $\mathbb{1}[.]$ is an indicator function.

3.6. Relation with the Expectation-Maximization

The proposed iterative approach can be reinterpreted as an Expectation-Maximization (EM) approach. To see this, consider a video $X = \{x_t\}_{t=1}^T$ containing T segments with the video label W. In our proposed approach, we first predict the segment labels Y and use them to compute the distribution of the number of positive segments Z. Here X, W are the observed variables, and Z is the latent variable. Therefore, we have the following graphical model:

$$P_\theta(X,Y,Z,W) = P(X)P_\theta(Y|X)P_\theta(Z|Y)P_\theta(W|Z) \tag{8}$$

where, $P_\theta(Y|X)$ is the segment classifier, and $P_\theta(W|Y,Z)$ is some MIL pooling operator. Then, to learn the model parameters θ from weakly-supervised data, we adopt an EM-based learning strategy. We alternate between the E-step and M-steps by optimizing for the evidence lower bound (ELBO) on the observed data log-likelihood $\log P(X,W)$ as,

$$\log P(X,W) \geq \mathbb{E}_{Q(Z|X,W)} \log \frac{p_\theta(X,Z,W)}{Q(Z|X,W)} \tag{9}$$

where, $Q(.|.)$ is any posterior distribution on latents. This is a result of Jensen's inequality and is tight if and only if $Q(Z|W,X)$ equals the true posterior $p(Z|X,W)$.

E-Step. The purpose of E-Step is to estimate the posterior on latent z given access to the latest model parameters θ'. For a given X, W, the latent variable can be estimated by

$$z^* = \arg\max_z P_{\theta'}(Z|X,W) \tag{10}$$

Here, we can further decompose the posterior using Eq. 8 as $P_{\theta'}(Z|X,W) = P(Z|Y)P_{\theta'}(Y|X,W)$. By plugging in our proposed Poisson binomial modeling (§3.3) for $P(Z|Y)$ and the pre-trained segment-classifier (§3.2) for $P(Y|X,W)$, we can arrive at our proposed approach of estimating dynamic threshold τ^* (Eq. 6) for latent z.

M-step. The M-step is to learn the model parameters θ by optimizing the ELBO from Eq. 9. By ignoring the terms that do not depend on model parameters, the objective of the M-step is,

$$\mathcal{L} \simeq \mathbb{E}_{Q(Z|X,W)} \log p_\theta(W|X,Z) \tag{11}$$

This is equivalent to optimizing the classification performance of the video classifier given z^*. Therefore, we can rewrite this as,

$$\mathcal{L} = \mathbf{CE}(p_\theta(W|X, Z = z^*), \mathbf{W}). \qquad (12)$$

In our approach, we implement the video classifier $p_\theta(W|X, Z = z^*)$ using the Poisson binomial-based MIL pooling (Eq. 4). Thus, our proposed loss (Eq. 5) is the M-step ELBO objective under weak-supervision constraints.

3.7. Relation with Previous MIL Methods

The proposed Poisson binomial distribution generalizes many of the existing MIL techniques for obtaining video-level probabilities from segment probabilities. The Noisy-OR (NOR) model [25, 48] estimates the bag-level probability as $\mathbf{P}(\sum_{t,m} y_i \geq 1) = 1 - \prod_{t,m}(1 - \hat{\mathbf{p}}_t^m)$. Using the Poisson binomial distribution, this is exactly $\sum_{k \geq 1} \hat{\mathbf{P}}_z(k; N)$. The max pooling based bag-level probability ($\hat{\mathbf{P}} = \max_{\forall t,m} \hat{\mathbf{p}}_t^m$) is included in $\hat{\mathbf{P}}_z(k = 1; N)$. The average-pooling based bag-level event probability ($\hat{\mathbf{P}} = \frac{1}{2T} \sum_{\forall t,m} \hat{\mathbf{p}}_t^m$) is equivalent to optimizing for the expected success of the distribution $\hat{\mathbf{P}}_z(k; N)$.

4. Experiments

4.1. Experimental setup

LLP dataset. We conduct our experiments on The *Look, Listen and Parse* (LLP) dataset [38] which consists of 11849 YouTube videos of 10 seconds duration, labelled into 25 event categories. These videos are unconstrained and consist of a wide variety of scene content including daily activities, music performances, vehicle sounds *etc*. We use 10,000 videos with weak labels (only video-level labels) for training. The rest of the 1,849 fully-annotated videos (with segment-level labels) are used for validation and testing. We use the standard train-val-test split from the dataset.
Evaluation Metrics. Following previous works [43, 38, 15], we use F1-scores on all types of events (audio, visual and audio-visual) as evaluation metrics. These metrics are computed both at the segment level and event level. To compute segment-level metrics, segment-level predictions are evaluated. Event level metrics are computed by computing F1-score on positive consecutive snippets in the same event with mIoU = 0.5 as the threshold. In addition, we also evaluate the overall audio-visual scene parsing performance of our method by computing aggregated results, i.e., "Type@AV" and "Event@AV". Specifically, Type@AV computes averaged audio, visual, and audio-visual event evaluation results, while Event@AV computes the F1-score considering all audio and visual events for each sample rather than directly averaging results from different events.
Implementation Details. For all experiments, we subsample the audio stream to 16 KHz and visual frames are processed at 8 fps. We use the same feature extractors as the baselines [43, 38, 15]. We use ResNet-152 [8] pretrained on ImageNet and R(2+1)D-18 [40] pretrained on Kinetics-400 as visual feature extractors to generate 512 dimensional feature. Audio features of 128-dimension are extracted from VGGish model [9] pretrained on AudioSet [6]. Our model is trained using Adam optimizer with a mini-batch of 16 and a learning rate of $3e-4$ for 20 epochs.

Table 1: Weakly-supervised audio-visual video parsing F1-score (%) comparison with different methods on the LLP dataset. Our proposed approach shows improvement over the baseline methods.

Methods	**Audio**		**Visual**		**Audio-Visual**		**Type@AV**		**Event@AV**	
	Seg.	**Event**	**Seg.**	**Event**	**Seg.**	**Event**	**Seg.**	**Event**	**Seg.**	**Event**
AVE [39]	47.2	40.4	37.1	34.7	35.4	31.6	39.9	35.5	41.6	36.5
AVSDN [19]	47.8	34.1	52.0	46.3	37.1	26.5	45.7	35.6	50.8	37.7
HAN [38]	60.1	51.3	52.9	48.9	48.9	43.0	54.0	47.7	55.4	48.0
Lamba *et al*. [15]	61.6	53.7	54.7	50.5	50.3	43.5	55.5	49.2	56.1	50.9
MA [43]	60.3	53.6	60.0	56.4	55.1	49.0	58.9	53.0	57.9	50.6
CVCM [20]	60.8	53.8	63.5	58.9	57.0	49.5	60.5	54.0	59.5	52.1
MGN-MA [26]	60.2	50.9	61.9	59.7	55.5	49.6	59.2	53.4	58.7	49.9
JoMoLD [3]	<u>61.3</u>	<u>53.9</u>	**63.8**	59.9	<u>57.2</u>	<u>49.6</u>	<u>60.8</u>	<u>54.5</u>	<u>59.9</u>	**52.5**
Ours	**63.1**	**54.1**	<u>63.5</u>	**60.3**	**57.7**	**51.5**	**61.4**	**55.2**	**60.6**	<u>52.3</u>

4.2. Quantitative and Qualitative Comparisons

We compare our method with the following baselines on weakly-supervised event detection methods - *AVE [39]* and *AVSDN [19]*. These approaches are designed for AVEL [39] task, which only focuses on identifying audio-visual events. To adopt these methods for AVVP, additional audio and visual branches are introduced [38]. We also compare with the following models proposed for weakly supervised AVVP task - *HAN [38]*, *Lamba* et al. *[15]*, *MA [43]*, *Lin* et al. *[20],* MGN [26], and JoMoLD [3]. These approaches use cross-modal and self-attention and train with MMIL setting along with additional losses. We refer to the results reported in [38, 43, 15, 3] for quantitative comparison. For qualitative comparison, we generate results by retraining the authors' publicly available code with default hyper-parameter settings.
Quantitative Results. For a fair comparison, all of these models are trained on the LLP dataset with the same data split. Table 1 shows the quantitative comparison of our method with the baseline methods. Our approach achieves an average improvement of 0.6 percentage points over the SOTA on the F1 score metric. This indicates that our method is able to localize and detect events more accurately, leading to an increase in both precision and recall. Given that the task under consideration is weakly supervised with no fine-grained supervision coupled with severe label imbalance for most of the events [15], this improvement is significant for this challenging task. We also show in §5 that our method is more robust and achieves more stable

Table 2: Analysis of runtime of our proposed Poisson Binomial pooling. N indicates the number of segments; Time in milliseconds.

N	10	20	50	80	100	200	300	500
Time	0.2	0.21	0.21	0.21	0.22	0.33	0.59	2.0

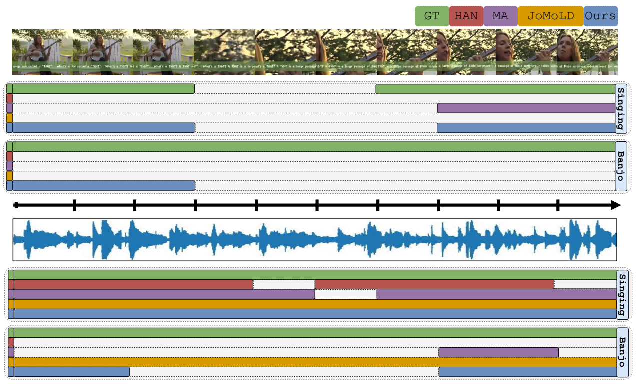

Figure 2: Audio-Visual Video Parsing results of our method with HAN [38], MA [43], JoMoLD [3] on one video.

results across different datasets and experimental settings.

Qualitative Analysis. We present some qualitative results in Figure 2. Here, we compare our method with HAN [38], MA [43] and JoMoLD [3] methods on one example and report the final parsing results. This video contains a musical scene with events "Singing" and "Banjo" occurring in both audio and visual modalities. While all three baseline methods (HAN, MA, JoMoLD) fail to localize "Banjo" in the visual modality, our method localizes when Banjo is completely visible ($0-2$ seconds). Even on the Visual-Singing event, our method localizes much better than MA, while HAN and JoMoLD fail to localize this event completely. In audio event localization, our method and JoMoLD perform better than MA and HAN by localizing the audio-singing event completely. On Audio-Banjo event, our method struggles to localize the event in its entirety when compared to JoMoLD. Overall, our proposed approach achieves better parsing results than baseline methods. These results indicate that our proposed approach of boosting positive segments in a video, using a model trained with weak-supervision constraints alone, helps in better scene parsing. We report a few more qualitative results in Supplementary.

On the convergence of number of positive segments. The parameter τ in Eq. 4 indicates the minimum number of positive segments (z) in a given video. Since we do not have any supervision on the number of positive segments, we iteratively estimate τ from a pre-trained model and then retrain the model using the proposed Poisson binomial based MIL loss. We empirically show that the model learns a better estimate of τ with the proposed iterative training. For this, we start with a model pre-trained with MA [43] setup and refine using our proposed approach. We report the histogram of the difference between the ground truth (τ^*) and the estimated count of total positive segments (τ) in a video on the test set for each stage of training in Figure 3. If the model learns a good estimate of τ, the term $\tau^* - \tau$ will have zero mean with a very small variance. From Figure 3, we can observe that a lot of the events are correctly identified ($\tau^* - \tau = 0$). We can observe that μ monotonically decreases (from 3.14 to 0.7 in visual modality; from 3.73 to 2.61 in audio modality) after each training epoch. The variance also reduces monotonically. This indicates that our approach, in fact, aids in capturing more positive segments.

Analysis of Runtime. The computational complexity of PoiBin (Eq. 3), which uses 1D-IDFT, is $O(N\log N)$, where N is the number of segments. We report the runtime of PoiBin-pooling (in milliseconds (ms)) for AVVP in Table 2. Given that HAN takes 5.9 ms (for $N = 10$), the computational overhead for implementing PoiBin (0.2 ms) is insignificant. Note that attentive-pooling takes 0.19 ms.

Table 3: Ablation studies to evaluate various components of the loss function.

Losses			Audio		Visual		AV		Type@AV		Event@AV	
$L_{\text{MIL}}^{\text{PoiBin}}$	$L_{\text{MIL}}^{\text{Att}}$	$L_{\text{MIL}}^{\text{aug}}$	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event	Seg.	Event
✓			61.4	51.9	55.9	52.2	53.1	46.3	56.8	56.9	50.1	48.4
	✓		60.0	52.9	60.1	56.3	54.7	46.8	58.2	52.6	57.4	50.6
✓	✓		61.1	53.3	62.6	58.9	57.7	51.5	60.8	54.6	59.1	51.5
✓		✓	60.6	52.5	58.4	54.6	54.0	46.6	57.6	54.3	54.4	49.7
	✓	✓	61.9	53.6	62.9	59.5	57.7	51.5	61.0	54.8	59.7	51.8
✓	✓	✓	**63.1**	**54.1**	**63.5**	**60.4**	**57.7**	**51.5**	**61.4**	**55.2**	**60.6**	**52.3**

4.3. Ablation study

We perform ablation studies to evaluate the effect of various components of our proposed approach.

Effect of losses: We perform an ablation study to analyze the impact of different components of our proposed objective (as defined in Eq. 3.5), and the results are reported in Table 3. The results show that our proposed Poisson binomial-based MIL formulation *without* augmentation alone (first row) is inferior to the existing Attention-based MIL formulation (second row) from MA [43]. But when our proposed loss is used in conjunction with MIL loss (third row), there is a significant improvement in performance. This suggests that our proposed loss function complements the existing method and can improve the overall performance. Incorporating data-augmentation, described in §3.5, along with these two losses (last row), the performance improves significantly. This is expected as the augmentation provides a better estimation of the latent variable (total number of positive segments in a video), leading to better temporal localization. This is expected as the proposed augmentation strategy helps in better localization by estimating positive segments. We can also observe that the proposed augmentation (fourth and fifth rows) improves

Table 4: Effect of τ on performance

	Audio		Visual		AV		Type@AV		Event@AV	
τ	**Seg.**	**Event.**	**Seg.**	**Event.**	**Seg.**	**Event.**	**Seg.**	**Event.**	**Seg.**	**Event.**
1	61.8	54.3	58.2	55.1	53.6	48.3	57.8	52.6	58.4	51.3
10	55.7	48.1	55.6	51.4	50.2	44.2	53.8	47.6	56	47.2
refine	63.1	54.1	63.4	59.9	57.7	51.5	61.4	55.2	60.6	52.3

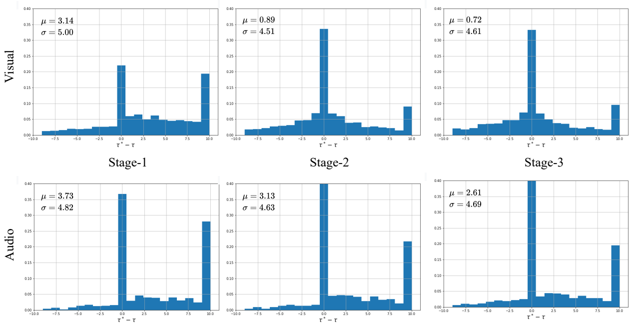

Figure 3: Effect of multi-stage training in estimating a better τ, which indicates a minimum number of positive segments in each video.

over the baselines without augmentation (first and second rows). Overall, our proposed approach complements the existing method and improves the performance of the model, indicating the effectiveness of our proposed formulation for a weakly-supervised task.

Effect of τ: We iteratively estimate τ from a pre-trained model, and then retrain the model using our proposed loss. We report results in Table 4. By fixing $\tau = 1$ (the lowest possible value, as $\tau = 0$ indicates the event does not occur), our algorithm performance is comparable to the baseline method. This is expected as our Poisson binomial formulation indicates that there is at least one positive segment in the video, which is exactly similar to the weak-supervision constraints. Therefore our formation is equivalent to the baseline model. A large τ implies that most of the segments contain the event, even when they may not. This would be detrimental to the training as such a setup is equivalent to training with many (noisy) false positive labels. Our experiments also show that this generally degrades the performance, as shown for the case when $\tau = 10$. On the other hand, in our proposed approach, we initialize τ to 1, which adheres to the weak-supervision constraints, and iteratively estimate a new τ as described in §3.4. Our experiments also show that this approach improves the performance (§4.2), and our model estimates better after each training epoch (Figure 3).

Usefulness of modeling $\mathbf{z}$ as latent over $\mathbf{y}$. Here, we perform ablation experiments by considering different choices for the latent variable in our formulation. Results reported in Table 5 indicate that our proposed modeling with z as

Table 5: Ablation studies to evaluate the modeling of z as the latent variable (+PL$_\mathbf{z}$) vs y as the latent variable (+PL$_\mathbf{y}$).

Methods	Audio		Visual		Audio-Visual		Type@AV		Event@AV	
	Seg.	**Event**	**Seg.**	**Event**	**Seg.**	**Event**	**Seg.**	**Event**	**Seg.**	**Event**
Baseline	60.0	52.9	60.1	56.3	54.7	46.8	58.2	52.6	57.4	50.6
+PL$_\mathbf{y}$	58.5	53.3	56.5	53.1	51.7	45.2	56.1	50.4	55.3	50.8
+PL$_\mathbf{z}$	**63.1**	**54.1**	**63.4**	**60.4**	**57.7**	**51.5**	**61.4**	**55.2**	**60.6**	**52.3**

Table 6: Results on Temporal Action Localization task on THUMOS14 dataset [11].

Method	IoU							AVG		
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**[0.1:0.5]**	**[0.3:0.7]**	**[0.1:0.7]**
Wang*et al.* [42]	44.4	37.7	28.2	21.1	13.7	-	-	20.6	-	-
W-TALC [30]	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8	-	-
EM-MIL [24]	59.1	52.7	45.5	36.8	30.5	22.7	16.4	44.9	30.4	37.7
Nguyen *et al.* [29]	60.4	56	46.6	37.5	26.8	17.6	9	45.5	27.5	36.3
HAM-Net [14]	65.4	59	50.3	41.1	31	20.7	11.1	49.4	30.9	39.8
FTCL [5]	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	34.4	43.6
UGCT [46]	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
DCC [16]	69.0	63.8	55.9	45.9	35.7	24.3	13.7	54.1	35.1	44.0
DGCNN [34]	66.3	59.9	52.3	43.2	32.8	22.1	13.1	50.9	32.7	41.3
Li *et al.* [17]	69.7	64.5	**58.1**	**49.9**	<u>39.6</u>	<u>27.3</u>	14.2	56.3	<u>37.8</u>	46.1
Huang *et al.* [10]	71.3	65.3	55.8	47.5	38.2	25.4	12.5	55.6	35.9	45.1
ASM-Loc [7]	71.2	65.5	<u>57.1</u>	46.8	36.6	25.2	13.4	55.4	35.8	45.1
DELU [2]	<u>71.5</u>	<u>66.2</u>	56.5	<u>47.7</u>	**40.5**	27.2	<u>15.3</u>	<u>56.5</u>	37.4	<u>46.4</u>
OURS	**72.7**	**66.4**	**58.1**	47.8	<u>39.6</u>	**28.6**	**19.2**	**56.9**	**38.6**	**47.5**

the latent variable is more robust than using segment-level pseudo-label y based formulation. These results support our initial hypothesis (in §3.3) that modeling the number of positive segments z as latent variables provides a more informative signal than weak labels while being less noisy than pseudo-segment labels ($\mathbf{y}$).

5. Generalization to other tasks

The proposed Poisson binomial formulation for the weakly supervised AVVP task does not make any task-specific assumptions. Therefore, we investigate the generalizability of our approach to other similar weakly supervised tasks. We perform preliminary experiments on the Temporal Action Localization (TAL) that aims to localize the start and end timestamps of action instances and recognize their categories simultaneously in untrimmed videos. We experiment with THUMOS14 dataset [11], which consists of videos with 100's of frames belonging to 20 action categories. We adopt Poisson binomial based MIL pooling to model this multi-class classification setup. The training details, along with architecture information, are available in Supplementary (§S3).

We report the results for this setup in Table 6. We evaluate in terms of mean Average Precision (mAP) with different temporal Intersection over Union (tIoU) thresholds, which is denoted as mAP@α where α is the threshold. Our model, trained with our proposed Poisson binomial-based MIL approach from §3.3, performs better than the current state-of-the-art model DELU [2].

Our model also shows more significant improvements at high threshold metrics tIoU=0.7, which implies that our action proposals are more complete. Note that, to use our proposed approach, we need to have a model that predicts per-frame probabilities only. Therefore, these results further indicate that our proposed approach has the potential to be further optimized and integrated with other MIL-based techniques to achieve even better results.

6. Limitations

Although our proposed approach performs better than state-of-the-art on AVVP and TAL tasks, there are a few limitations, which we discuss here. We can guarantee convergence by reinterpreting our proposed iterative approach as an EM algorithm. But it is known that EM algorithms converge to a locally optimal solution. From Table 1, performance gains in the audio-visual event (when an event co-occurs in both audio and visual modalities) are slightly lower than in audio events and visual events. One reason for this is that we are not modeling audio-visual events explicitly. We can potentially overcome this by modeling AV events, such as exploiting the temporal correlation between audio/visual modalities and using the sequential ordered nature of data. Learning better feature representations by employing explicit constraints on feature similarity may mitigate this issue. The LLP dataset suffers from severe label imbalance and strong label correlations [15], i.e., a set of labels co-occur more often than others. Our proposed strategy is not designed to address this issue.

7. Conclusions

In this paper, we proposed a method for improving temporal localization in a weakly-supervised audio-visual video parsing task. To this end, we proposed to model the total number of positive segments (z) in a video. We showed that this follows Poisson binomial distribution, which can be computed exactly from segment-level event probabilities. Since we do not have explicit supervision on the number of positive segments in a video, we proposed an iterative algorithm. We first estimate the minimum number of positive segments (τ) in a video and then optimize for the model parameters using the proposed Poisson binomial-based MIL loss. We also proposed a data-augmentation method that aided in improving the performance. Our proposed approach can be interpreted as an EM algorithm, which provides convergence guarantees. Experiments on the LLP dataset demonstrate that our proposed approach outperforms the state-of-the-art, validating the efficacy of our approach in improving the localization capacity under weak supervision. Additionally, our experiments on Temporal Action Localization demonstrate its potential for generalization to similar MIL tasks.

Acknowledgement

Support from Institute of Eminence (IoE) project No. SB22231269EEETWO005001 for Research Centre in Computer Vision is gratefully acknowledged. The first author thanks Lokesh Bommisetty for valuable discussions in the early stages of the work.

References

[1] Humam Alwassel, Fabian Caba Heilbron, Ali Thabet, and Bernard Ghanem. Refineloc: Iterative refinement for weakly-supervised action localization. 2019. 4

[2] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 192–208. Springer, 2022. 8

[3] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 431–448, Cham, 2022. Springer Nature Switzerland. 1, 6, 7

[4] Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. Temporal localization of actions with actoms. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2782–2795, 2013. 1, 2

[5] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19999–20009, 2022. 8

[6] 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. 6

[7] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13925–13935, June 2022. 8

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[9] 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 *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 6

[10] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3272–3281, 2022. 8

[11] 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. 8

[12] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 3

[13] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. *arXiv preprint arXiv:2101.00545*, 2021. 1, 2, 4

[14] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1637–1645, 2021. 8

[15] Jatin Lamba, Jayaprakash Akula, Rishabh Dabral, Preethi Jyothi, Ganesh Ramakrishnan, et al. Cross-modal learning for audio-visual video parsing. *arXiv preprint arXiv:2104.04598*, 2021. 1, 2, 4, 6, 9

[16] Jingjing Li, Tianyu Yang, Wei Ji, Jue Wang, and Li Cheng. Exploring denoised cross-video contrast for weakly-supervised temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19914–19924, 2022. 8

[17] Ziqiang Li, Yongxin Ge, Jiaruo Yu, and Zhongming Chen. Forcing the whole video as background: An adversarial learning strategy for weakly temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5371–5379, 2022. 8

[18] 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)*, pages 3–19, 2018. 1, 2

[19] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006. IEEE, 2019. 6

[20] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 6

[21] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 2

[22] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 344–353, 2019. 1, 2

[23] 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 *European conference on computer vision*, pages 729–745. Springer, 2020. 1, 2, 4

[24] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 729–745. Springer, 2020. 8

[25] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 6

[26] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 4, 6

[27] Sanath Narayan, Hisham Cholakkal, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13608–13617, 2021. 2

[28] 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*, pages 6752–6761, 2018. 1, 2, 4

[29] 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*, pages 5502–5511, 2019. 8

[30] 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)*, pages 563–579, 2018. 8

[31] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4372–4376, 2020. 2

[32] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 2

[33] Baifeng Shi, Qi Dai, Yadong Mu, and Jingdong Wang. Weakly-supervised action localization by generative attention modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1009–1019, 2020. 1, 2, 4

[34] Haichao Shi, Xiao-Yu Zhang, Changsheng Li, Lixing Gong, Yong Li, and Yongjun Bao. Dynamic graph modeling for

weakly-supervised temporal action localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3820–3828, 2022. 8
[35] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5734–5743, 2017. 1, 2
[36] 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*, pages 1049–1058, 2016. 1, 2
[37] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *2017 IEEE international conference on computer vision (ICCV)*, pages 3544–3553. IEEE, 2017. 2
[38] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. 1, 2, 3, 4, 6, 7
[39] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 1, 2, 6
[40] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6
[41] 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*, pages 4325–4334, 2017. 1, 2, 4
[42] 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*, pages 4325–4334, 2017. 8
[43] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 1, 2, 4, 6, 7
[44] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3893–3901, 2020. 2
[45] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–63, 2021. 2
[46] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–63, 2021. 8
[47] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *European conference on computer vision*, pages 37–54. Springer, 2020. 2, 4
[48] Cha Zhang, John Platt, and Paul Viola. Multiple instance boosting for object detection. *Advances in neural information processing systems*, 18, 2005. 6
[49] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *European Conference on Computer Vision*, pages 539–555. Springer, 2020. 1, 2
[50] 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*, pages 2914–2923, 2017. 1, 2
[51] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 35–44, 2018. 2
[52] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8436–8444, June 2021. 2