

Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning

Zhekun Luo¹ Devin Guillory¹ Baifeng Shi² Wei Ke³ Fang Wan⁴
 Trevor Darrell¹ Huijuan Xu¹

¹University of California, Berkeley ²Peking University

³Carnegie Mellon University ⁴Chinese Academy of Sciences

¹{zhekun_luo, dguillory, trevordarrell, huijuan}@eecs.berkeley.edu,
²bfshi@pku.edu.cn, ³weik@andrew.cmu.edu, ⁴wanfang@ucas.ac.cn

Abstract. Weakly-supervised action localization requires training a model to localize the action segments in the video given only video level action label. It can be solved under the Multiple Instance Learning (MIL) framework, where a bag (video) contains multiple instances (action segments). Since only the bag’s label is known, the main challenge is assigning which key instances within the bag to trigger the bag’s label. Most previous models use attention-based approaches applying attentions to generate the bag’s representation from instances, and then train it via the bag’s classification. These models, however, implicitly violate the MIL assumption that instances in negative bags should be uniformly negative. In this work, we explicitly model the key instances assignment as a hidden variable and adopt an Expectation-Maximization (EM) framework. We derive two pseudo-label generation schemes to model the E and M process and iteratively optimize the likelihood lower bound. We show that our EM-MIL approach more accurately models both the learning objective and the MIL assumptions. It achieves state-of-the-art performance on two standard benchmarks, THUMOS14 and ActivityNet1.2.

Keywords: weakly-supervised learning, action localization, multiple instance learning

1 Introduction

As the growth of video content accelerates, it becomes increasingly necessary to improve video understanding ability with less annotation effort. Since videos can contain a large number of frames, the cost of identifying the exact start and end frames of each action is high (frame-level) in comparison to just labeling what actions the video contains (video-level). Researchers are motivated to explore approaches that do not require per-frame annotations. In this work, we focus on weakly-supervised action localization paradigm, using only video-level action labels to learn activity recognition and localization. This problem can be framed as a special case of the **Multiple Instance Learning (MIL)** problem [4]: a

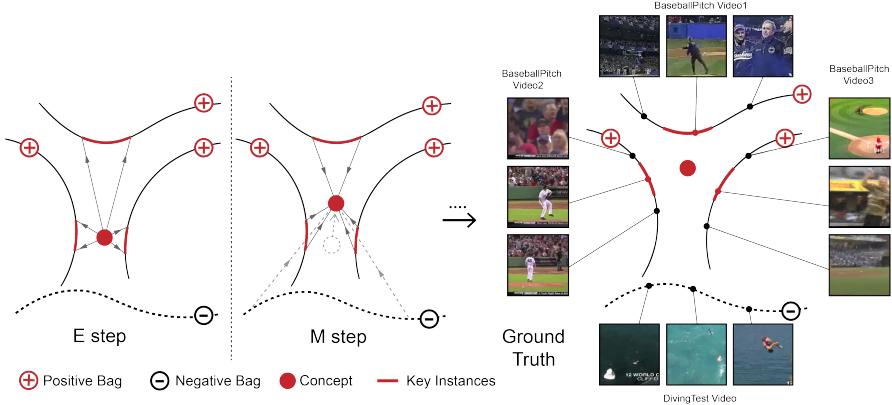


Fig. 1: Each curve represents a bag and points on the curve represent instances in the bag. We aim to find a concept point such that each positive bag contains some key instances close to it while all instances in the negative bags are far from it. In E step we use the current concept to pick key instances for each positive bag. In M step we use key instances and negative bags to update the concept.

bag contains multiple instances; Instances' labels collectively generate the bag's label, and only the bag's label is available during training. In our task, each video represents bag, and the clips of the video represent the instances inside the bag. The key challenge here is to handle **key instance assignment** during training – to identify which instances within the bag trigger the bag's label.

Most previous works used **attention-based** approaches to model the key instance assignment process. They used attention weights to combine instance-level classification to produce the bag's classification. Models of this form are then trained via standard classification procedures. The learned attention weights imply the contribution of each instance to the bag's label, and thus can be used to localize the positive instances (action clips) [17,26]. While promising results have been observed, models of this variety tend to produce incomplete action proposals [13,31], that only part of the action is detected. This is also a common problem in attention-based weakly-supervised object detection [11,25]. We argue that this problem is due to a misspecification of the MIL-objective. Attention weights, which indicate key instances' assignment, should be our optimization target. But in an attention-MIL framework, attention is learned as a by-product when conducting classification for bags. As a result, the attention module tends to only pick the most discriminative parts of the action or object to correctly classify a bag, due to the fact that the loss and training signal come from the bag's classification.

Inspired by traditional MIL literature, we adopt a different method to tackle weakly-supervised action localization using the ExpectationMaximization framework. Historically, ExpectationMaximization (EM) or similar iterative estimation processes have been used to solve the MIL problems [4,5,35] before the deep

learning era. Motivated by these works, we explicitly model key instance assignment as a hidden variable and optimize this as our target. Shown in Fig. 1, we adopt the EM algorithm to solve the interlocking steps of key instance assignment and action concept classification. To formulate our learning objective, we derive two pseudo-label generating schemes to model the E and M process respectively. We show that our alternating update process optimizes a lower bound of the MIL-objective. We also find that previous attention-MIL models implicitly violate the MIL assumptions. They apply attention to negative bags, while the MIL assumption states that instances in negative bags are uniformly negative. We show that our method can better model the data generating procedure of both positive and negative bags. It achieves state-of-the-art performance with a simple architecture, suggesting its potential to be extended to many practical settings. The main contributions of this paper are:

- We propose to adapt the ExpectationMaximization MIL framework to weakly supervised action localization task. We derive two novel pseudo-label generating schemes to model the E and M process respectively.¹
- We show that previous attention-MIL models implicitly violate the MIL assumptions, and our method better model the background information.
- Our model is evaluated on two standard benchmarks, THUMOS14 and ActivityNet1.2, and achieves state of the art results.

2 Related Work

Weakly-Supervised Action Localization Weakly supervised action localization learns to localize activities inside videos when only action class labels are available. UntrimmedNet [26] first used attention to model the contribution of each clip to a video-level action label. It performs classification separately at each clip, and predicts video’s label through a weighted combination of clips’ scores. Later the STPN model [17] proposed that instead of combining clips’ scores, it uses attention to combine clips’ features into a video-level feature vector and conducts classification from there. [8] generalizes a framework for these attention-based approaches and formalizes such combination as a permutation-invariant aggregation function. W-TALC [19] proposed a regularization to enforce action periods of the same class must share similar features. It is also noticed that attention-MIL methods tend to produce incomplete localization results. To tackle that, a series of papers [22,23,33,38] took the adversarial erasing idea to improve the detection completeness by hiding the most discriminative parts. [31] conducted sub-samplings based on activation to suppress the dominant response of the discriminative action parts. To model complete actions, [13] proposed to use a multi-branch network with each branch handling distinctive action parts. To generate action proposals, they combine per-clip attention and classification scores to form the Temporal Class Activation Sequence (T-CAS

¹ Code: <https://github.com/airmachine/EM-MIL-WeaklyActionDetection>

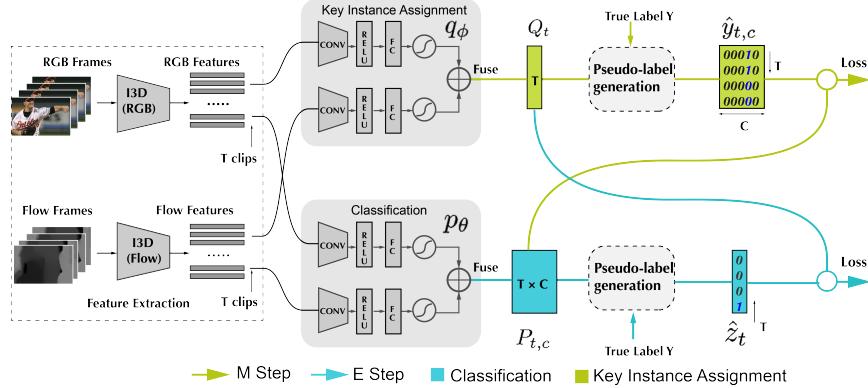


Fig. 2: Our EM-MIL model architecture builds on fixed two-stream I3D features, and alternates between updating the key-instance assignment branch q_ϕ (E Step) and the classification branch p_θ (M Step). We use the classification score and key instance assignment result to generate pseudo-labels for each other (detailed in Sec. 3.1 and Sec. 3.2), and alternate freezing one branch to train the other.

[17]) and group the high activation clips. Another type of models [21,14] train a boundary predictor based on pre-trained T-CAS scores to output the action start and end point without grouping.

Some previous methods in weakly-supervised object or action localization involve iterative refinement, but their training processes and objectives are different from our ExpectationMaximization method. RefineLoc [1]'s training contains several passes. It uses the result of the i^{th} pass as supervision for the $(i + 1)^{th}$ pass and trains a new model from scratch iteratively. [24] uses a similar approach in image objection detection but stacks all passes together. Our approach differs from these in the following ways: Their self-supervision and iterative refinement happen between **different passes**. In each pass all modules are trained jointly till converge. In comparison, we adopts an EM framework which explicitly models key instance assignment as hidden variables. Our pseudo-labeling and alternating training happen between **different modules** of the same model. Thus our model requires only one pass. In addition, as discussed in Sec. 3.4, they handle the attention in negative bags different to us.

Traditional Multi-Instance Learning Methods The Multiple Instance Learning problem was first defined by Dietterich et al. [4], who proposed the iterated discrimination algorithm. It starts from a point in the feature space and iteratively searches for the smallest box covering at least one point (instance) per positive bag and avoiding all points in negative bags. [15] sets up the Diverse Density framework. They defined a point in the feature space to be the positive

concept. Every positive bag (“diverse”) contains at least one instance close to the concept while all instances in the negative bags are far from it (in terms of some distance metric). They modeled the likelihood of a concept using Gaussian Mixture models along with a Noisy-OR probability estimation. [34] then applied AdaBoost to this Noisy-OR model and [10]’s ISR model, and derived two MIL loss functions. [5] adapted the K-nearest neighbors method to the Diverse Density framework. Later [35] proposed the EM-DD algorithm, combining Expectation Maximization process and the Diverse Density metric. These early works did not involve neural networks and were not applied over the high-dimensional task of action localization. Many of them involve modeling key instances assignment as hidden variable and use iterative optimization. They also differ from the predominant attention-MIL paradigm in how they treat negative instances. We view these distinctions as motivation to explore our approach.

3 Method

Multiple Instance Learning (MIL) is a supervised learning problem where instead of one instance X being matched to one label y , a bag or set of multiple instances $[X_1, X_2, X_3, \dots]$ are matched to single label y . In the binary MIL setting, a bag’s label is positive if at least one instance in the bag is positive. Therefore a bag is negative only if all instances in the bag are negative.

In our task, following the best practice of previous works [17,19,26], we divide a long video into multiple 15-frame clips. Then a video corresponds to a bag (bag-level video label is given), and the clips of the video represent the instances inside the bag (instance-level clip labels are missing). Each video (bag) contains T video clips (instances), denoted by $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T$, where $\mathbf{x}_t \in \mathbb{R}^d$ is the feature of clip t . We represent the video’s action label in one hot way, where $y_c = 1$ if the video contains clips of action c , otherwise $y_c = 0$, $c \in \{1, 2, \dots, C\}$ (each video can contain multiple action classes). In the MIL setting, label of each video is determined by the labels of clips it contains. To be specific, we assign a binary variable $z_t \in \{0, 1\}$ to each clip t , denoting whether clip t is responsible for the generation of video-level label. $\mathbf{z} = \{z_t\}_{t=1}^T$ models the assignment of **key instances scope**. Video-level label is generated with probability:

$$p_\theta(y_c = 1 | \mathbf{X}, \mathbf{z}) = \sigma_{t \in \{1, \dots, T\}} \{ p_\theta(y_{c,t} = 1 | \mathbf{x}_t) \cdot [z_t = 1] \} \quad (1)$$

where $[z_t = 1]$ is the indicator function for assignment. $p_\theta(y_{c,t} = 1 | \mathbf{x}_t)$ is the probability (parameterized by θ) that clip t belongs to class c . The closer clip t is to the concept, the higher $p_\theta(y_{c,t} = 1 | \mathbf{x}_t)$ is. σ is a permutation-invariant operator, e.g. maximum [36] or mean operator [8].

In our temporal action localization problem, we propose to first estimate the probability of $z_t = 1$ with an estimator $q_\phi(z_t = 1 | \mathbf{x}_t)$ parameterized by ϕ , and then choose the clips with high estimated likelihood as our action segments. Since $\{z_t\}$ are latent variables with no ground truth, we optimize q_ϕ through

maximization of the variational lower bound:

$$\begin{aligned} \log p_\theta(y_c|\mathbf{X}) &= KL(q_\phi(\mathbf{z}|\mathbf{X}) \parallel p_\theta(\mathbf{z}|\mathbf{X}, y_c)) + \int q_\phi(\mathbf{z}|\mathbf{X}) \log \frac{p_\theta(\mathbf{z}, y_c|\mathbf{X})}{q_\phi(\mathbf{z}|\mathbf{X})} d\mathbf{z} \\ &\geq \int q_\phi(\mathbf{z}|\mathbf{X}) \log p_\theta(\mathbf{z}, y_c|\mathbf{X}) d\mathbf{z} + H(q_\phi(\mathbf{z}|\mathbf{X})), \end{aligned} \quad (2)$$

where $H(q_\phi(\mathbf{z}|\mathbf{X}))$ is entropy of q_ϕ . By maximizing the lower bound, we are actually optimizing the likelihood of y_c given \mathbf{X} . In this work, we adopt the Expectation-Maximization (EM) algorithm, and optimize the lower bound by updating θ and ϕ alternately. To be specific, we first update ϕ by minimizing $KL(q_\phi(\mathbf{z}|\mathbf{X}) \parallel p_\theta(\mathbf{z}|\mathbf{X}, y_c))$ and tighten the lower bound in E step, and update θ through maximization of the lower bound in M step. In the following subsections, we will first get into details of updating θ and ϕ in E step and M step separately, and then sum up the whole algorithm.

3.1 E Step

In E step, we update ϕ by minimizing $KL(q_\phi(\mathbf{z}|\mathbf{X}) \parallel p_\theta(\mathbf{z}|\mathbf{X}, y_c))$ and tighten the lower bound in Eq. 2. As in previous works [17, 18], we approximate $q_\phi(\mathbf{z}|\mathbf{X})$ with $\prod_t q_\phi(z_t|\mathbf{x}_t)$ assuming the independence between different clips, where $q_\phi(z_t|\mathbf{x}_t)$ is estimated by neural network with parameter ϕ on each clip. Thus we only have to minimize $KL(q_\phi(z_t|\mathbf{x}_t) \parallel p_\theta(z_t|\mathbf{x}_t, y_c))$ for each clip t . Following the literature, we assume that the posterior $p_\theta(z_t|\mathbf{x}_t, y_c)$ is proportional to the classification score $p_\theta(y_c|\mathbf{x}_t)$. Then we propose to update q_ϕ with pseudo label generated from classification score. Specifically, dynamic thresholds are calculated based on the instance classification scores to generate pseudo-labels for q_ϕ . If an instance has a classification score over the threshold for any ground truth class within the video, the instance is treated as a positive example; otherwise, it is treated as a negative example. The pseudo label is formulated as follows:

$$\hat{z}_t = \begin{cases} 1, & \text{if } \sum_{c=1}^C \mathbb{1}(P_{t,c} > \bar{P}_{1:T,c} \wedge y_c = 1) > 0 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $P_{t,c} = p_\theta(y_c|\mathbf{x}_t)$ and $\bar{P}_{1:T,c}$ is the mean of $P_{t,c}$ over temporal axis. Then we update q_ϕ using binary cross entropy (BCE) loss and the updating process is illustrated in Fig. 3.

$$\mathcal{L}(q_\phi) = -\hat{z}_t \log q_\phi(z_t|\mathbf{x}_t) - (1 - \hat{z}_t) \log(1 - q_\phi(z_t|\mathbf{x}_t)). \quad (4)$$

3.2 M Step

In M step, we update p_θ through optimization of the lower bound in Eq. 2. Since $H(q_\phi(\mathbf{z}|\mathbf{X}))$ is constant wrt θ , we only optimize $\int q_\phi(\mathbf{z}|\mathbf{X}) \log p_\theta(\mathbf{z}, y_c|\mathbf{X}) d\mathbf{z}$, which is equivalent to optimize the classification performance given key instance assignment $q_\phi(\mathbf{z}|\mathbf{X})$. To this end, we use the class-agnostic key-instance assigning

module q_ϕ and the ground truth video-level labels to generate a $T \times C$ pseudo-label map which discriminates between foreground and background clips within the same video. Similarly, our pseudo-label generation procedure calculates a dynamic threshold based on the distribution of instance-assignment scores for each video clip. It assigns positive classifications for all instances whose scores are higher than the threshold, and negative classifications for all instances whose scores are below or instances in negative bags. The pseudo label is given by:

$$\hat{y}_{t,c} = \begin{cases} 1, & \text{if } y_c = 1 \text{ and } Q_t > \bar{Q}_{1:T} + \gamma \cdot (\max(Q_t) - \min(Q_t)) \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where $Q_t = q_\phi(z_t | \mathbf{x}_t)$ and $\bar{Q}_{1:T}$ is the mean of Q_t over temporal axis. The threshold hyper-parameter γ implies a distribution priori on how similar the same action exhibits across several videos. Then we update p_θ with BCE loss and the updating process is illustrated in Fig. 4.

$$\mathcal{L}(p_\theta) = -\hat{y}_{t,c} \log p_\theta(y_c | x_t) - (1 - \hat{y}_{t,c}) \log(1 - p_\theta(y_c | x_t)). \quad (6)$$

3.3 Overall Algorithm

We summarize our EM-style algorithm in Alg. 1. We update the key-instance assigning module q_ϕ and classification module p_θ alternately. In E step we freeze the classification p_θ and update q_ϕ using pseudo labels from p_θ . In M step we optimize classification based on q_ϕ . Two steps are processed alternately to maximize the likelihood $\log p_\theta(y_c | \mathbf{X})$, and meanwhile optimize the localization results.

Algorithm 1: EM-MIL Weakly-Supervised Activity Localization

Initialization: learning rate β , classification threshold γ
 classifier parameters θ , attention parameters ϕ

```

while  $\theta, \phi$  has not converged do
    #Estep
    for ( $\mathbf{X}, y_c$ ) in train set do
         $P_{t,c} \leftarrow p_\theta(y_c | \mathbf{x}_t)$  ;
         $\phi \leftarrow \phi - \beta \cdot \nabla_\phi \mathcal{L}(q_\phi)$  ;
    end
    #Mstep
    for ( $\mathbf{X}, y_c$ ) in train set do
         $Q_t \leftarrow q_\phi(z_t | \mathbf{x}_t)$  ;
         $\theta \leftarrow \theta - \beta \cdot \nabla_\theta \mathcal{L}(p_\theta)$  ;
    end
end

```

3.4 Comparison with Previous Methods

After careful examination of Eq. 3 and Eq. 5, we find that our pseudo-labeling process Q_t and $\hat{y}_{t,c}$ can also be interpreted as a special kind of attention. Denote

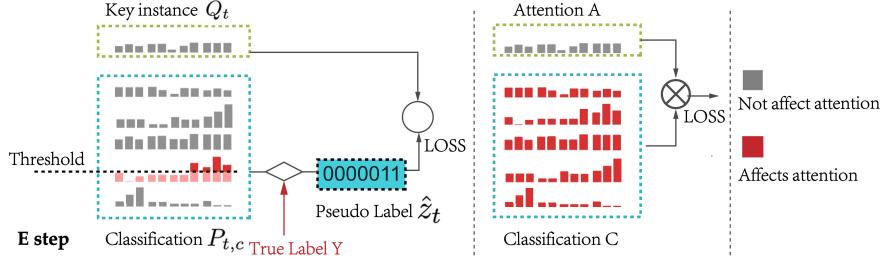


Fig. 3: In our EM-MIL model only the foreground classification score $P_{t,c}$ affects the key instance pseudo label \hat{z}_t (left), while in previous models all-class classification scores contribute to the attention weights (right).

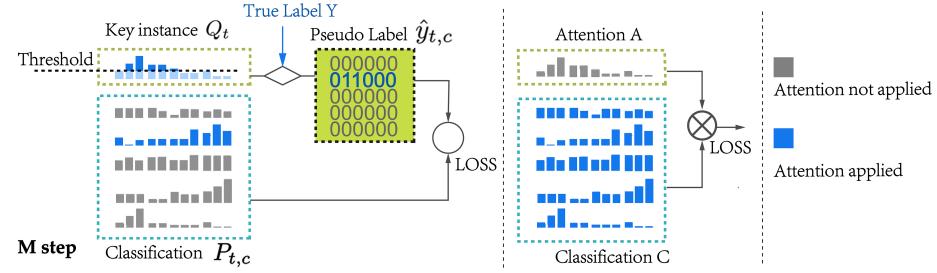


Fig. 4: Our EM-MIL model (left) uses key instance assignment Q_t to generate pseudo classification labels $\hat{y}_{t,c}$ only for the foreground classes, while in previous models such as UntrimmedNet (right) attentions are applied to all classes.

loss function by \mathcal{L} , then in Eq. 5, the loss is calculated as

$$\mathcal{L} [p_\theta(\mathbf{y}|\mathbf{x}), \mathcal{F}(\mathbf{Q}, \mathbf{y})] \quad (7)$$

\mathcal{F} is the pseudo label generation function in Eq. 5, $\mathbf{Q}, \mathbf{y}, \mathbf{x}$ is the compact expression of Q_t, y_c, x_t . On the other hand, if we denote attention and classification score as \mathbf{a}, \mathbf{c} , the loss for a typical attention-based model like [26] is:

$$\mathcal{L} [\sigma(\mathbf{c} \odot \mathbf{a}), \mathbf{y}] \quad (8)$$

Here σ is the aggregation operator [8], such as *reduce_sum* or *reduce_max*. Comparing Eq. 7 to Eq. 8, it is easy to see that they can be matched. $p_\theta(\mathbf{y}|\mathbf{x})$ is classification score (\mathbf{c}), and \mathbf{Q} can be seen as special attention (corresponds to \mathbf{a}). In M step it attends to the key instance it estimates. But compared to previous attention-MIL methods, Eq. 3 shows that this “attention” only happens in positive bags. We believe it better aligns with the MIL assumption, which says that all instances in negative bags are uniformly negative. Previous methods that applies attention to negative bags implicitly assumes that some instances are more negative than others. This violates the MIL assumption. The differences between our attention and theirs are illustrated in Fig. 3 and 4. In addition, in Eq. 5,

this ‘‘attention’’ is a threshold-based hard attention. Clips below the threshold are classified as background with high confidence, while clips above the threshold are weighted equally and re-scored in the next iteration. The use of hard pseudo labels allows for the distinct treatment of positive and negative instances that would be more complex to enforce with soft-boundaries. We initialize our training procedure by labeling every clip in a positive bag to be 1 and gradually narrow down the search scope. Such training process maintains high recalls for action clips in each E-M iteration. It prevents attention from focusing on the discriminative parts too quickly, thus increases the proposal completeness.

Another way to compare our methods with previous ones is through the lens of the MIL framework. As discussed in [2], MIL problem has two setting: instance-level vs bag-level. The instance level setting prioritizes classification precision of instance over bag’s, and vice versa. Our task aligns with the instance setting as the primary goal is action localization (equivalent to clips’ classification). Previous attention-MIL models like [17,19,26] treat instance-localization as the by-product of an accurate bag-level classification system, which align with the bag-level MIL setting. By modeling the problem through an instance-level MIL framework our approach more accurately models the target objective. This change in objective function and optimization procedure allows substantial improvement in performance.

3.5 Inference

At test time, we use another branch for video-level classification and use our model for localization as in previous work [21]. For classification branch, we used a plain UntrimmedNet [26] with soft attention for the THUMOS14 dataset and the W-TALC [19] for the ActivityNet1.2 dataset. We run a forward pass with our model to get the localization score L by fusing instance assignment score Q_t and classification score $P_{t,c}$.

$$L_t = \lambda * Q_t + (1 - \lambda) * P_{t,c}, \quad (9)$$

where λ is set to be 0.8 through grid search in THUMOS14 dataset and 0.3 in the ActivityNet1.2 dataset. In the Experiment Sec. 4.2 we analyze the impact of different of λ . We threshold the L_t score to get prediction y'_i for each clip using the same scheme as in Eq. 5. Then we group the clips above the threshold to get the temporal start and end point of the action proposal.

4 Experiments

In this section, we evaluate our EM-MIL model on two large-scale temporal activity detection datasets: THUMOS14 [9] and ActivityNet1.2 [7]. Sec. 4.1 introduces experimental setup of these datasets, the evaluation metrics and the implementation details. Sec. 4.2 compares weakly localization results between our proposed model and the state-of-the-art models on both THUMOS14 and ActivityNet1.2 datasets, and visualizes some localization results. Sec. 4.3 shows the ablation studies for each component of our model on THUMOS14 dataset.

4.1 Experimental Setup

Datasets: The THUMOS14 [9] activity detection dataset contains over 24 hours of videos from 20 different athletic activities. The train set contains 2765 trimmed videos, while the validation set and the test set contains 200 and 213 untrimmed videos respectively. We use the validation set as train data and report weakly-supervised temporal activity localization results on the test set. This dataset is particularly challenging as it consists of very long videos with multiple activity instances of very small duration. Most videos contain multiple activity instances of the same activity class. In addition, some videos contain activity instances from different classes.

The ActivityNet [7] dataset consists three versions. We use the ActivityNet1.2 version which contains a total of around 10000 videos including 4819 train videos, 2383 validation videos, and 2480 withheld test videos for challenge purpose. We report the weakly-supervised temporal activity localization results on the validation videos. In ActivityNet1.2, around 99% videos contain activity instances of a single class. Many of the videos have activity instances covering more than half of the duration. Compared to THUMOS14, this is a large-scale dataset, both in terms of the number of activities involved and the amount of videos.

Evaluation Metric: The weakly-supervised temporal activity localization results are evaluated 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. Average mAP at 10 evenly distributed tIoU thresholds between 0.5 and 0.95 is also commonly used in the literature.

Implementation Details: Video frames are sampled at 12 fps (for THUMOS14) or 25 fps (for ActivityNet1.2). For each frame, we perform the center crop of size 224×224 after re-scaling the shorter dimension to 256 and construct video clips for every 15 frames. We extract the features of the clips using the publicly released, two-stream I3D model pretrained on Kinetics dataset [3]. We use the feature map from *Mixed_5c* layer as feature representation. For optical flow stream, TV-L1 flow [27,32] is used as the input.

Our model is implemented in pyTorch and trained using Adam optimizer with initial learning rate 0.0001 for both datasets. For the THUMOS14 dataset, we train the model by alternating E/M step every 10 epochs in the first 30 epochs. Then we raise the learning rate to 4 times larger and decrease the alternating cycle to 1 epoch for another 35 epochs. For ActivityNet1.2 dataset, we use a similar training approach but the alternating cycle is 5 epochs and the learning rate is constant. We use our model to generate instance assignment Q_t and classification score $P_{t,c}$ separately for RGB and Flow branch. Then, we fuse the RGB/Flow score by weighted averaging. The threshold hyper-parameter γ in Eq. 5 is set to 0.15 for THUMOS14 dataset and 0 for ActivityNet1.2 dataset. Intuitively, the value of γ reflects how similar the same action exhibits across several videos, and should be negatively correlated with the variance of the action’s feature distribution. We also explore different γ in the range of [0.05,

Table 1: Our EM-MIL detection results on THUMOS14 in percentage. mAP at different tIoU thresholds α are reported. The top half shows fully-supervised methods while the bottom half shows weakly-supervised ones including ours. EM-MIL-UNT represents the result using UntrimmedNet’s [26] features.

Supervision	Models	α						
		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Fully-Supervised	CDC [20]	-	-	40.1	29.4	23.3	13.1	7.9
	R-C3D [28]	54.5	51.5	44.8	35.6	28.9	-	-
	Gao et al. [6]	-	-	50.1	41.3	31.0	19.1	9.9
	SSN [37]	66.0	59.4	51.9	41.0	29.8	19.6	10.7
	Xu et al. [29]	56.9	54.7	51.2	43.0	36.1	-	-
	BSN [12]	-	-	53.5	45.0	36.9	28.4	20.0
Weakly-Supervised	Hide [22]	36.4	27.8	19.5	12.7	6.8	-	-
	UntrimmedNet [26]	44.4	37.7	28.2	21.1	13.7	-	-
	STPN [17]	52.0	44.7	35.5	25.8	16.9	9.9	4.3
	Autoloc [21]	-	-	35.8	29.0	21.2	13.4	5.8
	W-TALC [19]	55.2	49.6	40.1	31.1	22.8	-	7.6
	RefineLoc-I3D [1]	-	-	40.8	-	23.1	-	5.3
	Liu et al. [14]	-	-	37.0	30.9	23.9	13.9	7.1
	Yu et al. [30]	-	-	39.5	-	24.5	-	7.1
	3C-Net [16]	59.1	53.5	44.2	34.1	26.6	-	8.1
	Nguyen et al. [18]	64.2	59.5	49.1	38.4	27.5	17.3	8.6
	EM-MIL (ours)	59.1	52.7	45.5	36.8	30.5	22.7	16.4
	EM-MIL-UNT (ours)	59.0	50.4	42.7	34.5	27.2	18.9	10.2

0.2], mAP@tIoU=0.5 varies between 29.0% and 30.5% in THUMOS14 dataset, compared to the previous SOTA 26.8% [18] using the same training data.

4.2 Comparison with State-of-the-art Approaches

Results on THUMOS14 Dataset: We compare our model’s results on the THUMOS14 dataset with state-of-the-art results in Table 1. Our model outperforms all the previous published models and achieves a new state-of-the-art result at mAP@0.5, **30.5%**. This result is achieved by our simple EM training policy and the pseudo-labeling scheme, without auxiliary losses to regularize the learning process. Compared to the best result among the six recent models [1,16,17,18,19,30] using the same two-stream I3D feature extraction backbone as our model, we get 3% significant improvement at mAP@0.5. We also tried using UntrimmedNet’s feature on our model (denoted as EM-MIL-UNT in Table 1), and got a mAP@0.5 of 27.2% which still improves significantly over previous models (e.g. [14,21,26]) using the same feature backbone. Our model also shows more significant improvement at high threshold metrics tIoU=0.6 and tIoU=0.7, which implies that our action proposals are more complete. On the other hand, our performance is slightly worse in the low IoU metrics.

Several examples’ qualitative results are shown in Fig. 5(a). For each example, we show the video, intermediate score map L_t from our model, final activity

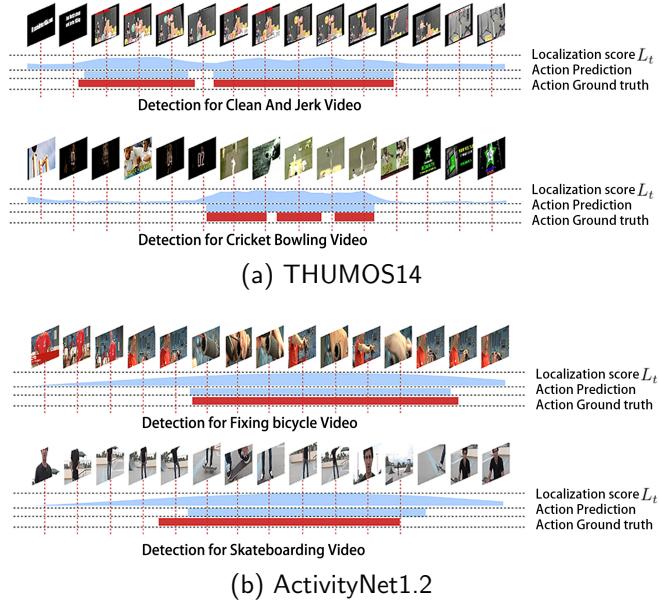


Fig. 5: Qualitative visualization. (a) and (b) show results for two videos each on THUMOS14 and ActivityNet1.2, a good prediction example (top) and a bad one (bottom). Ground truth activity segments are marked in red. Localization score distribution L_t and predicted activity segments are in blue.

detection result and ground truth temporal segment annotation. In the first example of *Clean and Jerk*, we localize the activity correctly with almost 100% overlap. We also show one bad prediction from our model in the second example, where our model overestimates the *Cricket Bowling* activity duration by 20%, as an effect of the interactive shrinkage training process which first labels every instance positive. Our model greatly resolves the incompleteness problem for activity detection in videos containing multiple action segments, while in some cases it might also bring in additional false positives. In addition, our model is also highly time efficient: in THUMOS14 our model trains for 65 epochs, taking 64.7s on two TITAN RTX GPUs. We have run the released code for AutoLoc [21] and W-TALC [19] on the same machine with their recommended training procedures. Their training times are 44.5s and 6051.2s, respectively. All experiments used pre-computed features and [21]'s training required additional pretrained CAS scores.

Results on ActivityNet1.2 Dataset: We compare our model's results on the ActivityNet1.2 dataset with previous results in Table 2. Our model outperforms previously published models in mAP@0.5 and gets the value of **37.4%**. Despite the state-of-the-art result in mAP@0.5, our model performs worse in high tIoU metrics, which is the opposite to what we observed on THUMOS14

Table 2: Detection results on ActivityNet1.2 in terms of mAP@{0.5, 0.7, 0.9} and average mAP at tIoU thresholds $\alpha \in (0.5, 0.95)$ with step 0.05 (in percentage). It shows both fully-supervised method and weakly-supervised ones.

Supervision	Models	α			avg. mAP
		0.5	0.7	0.9	
Fully-Supervised	SSN [37]	41.3	30.4	13.2	26.6
	UntrimmedNet [26]	7.4	3.9	1.2	3.6
	Autoloc [21]	27.3	17.5	6.8	16.0
	W-TALC [19]	37.0	14.6	4.2	18.0
	3C-Net [16]	37.2	23.7	9.2	21.7
	Liu et al. [14]	37.1	23.4	9.2	21.6
	TSM [30]	28.3	18.9	7.5	17.1
EM-MIL (ours)		37.4	23.1	2.0	20.3

dataset. We further investigate the reason for different result trends on both datasets. Videos in the THUMOS14 dataset contains multiple action segments, each segment with relatively short duration. It has high localization requirement where our model outperforms previous ones at high tIoU. Unlike THUMOS14, most videos ($> 99\%$) in the ActivityNet1.2 dataset have only one action class, and most of these videos have only a few activity segments which compose a big portion of the whole video duration. Thus videos in ActivityNet1.2 dataset can be regarded as trimmed actions in certain extent. We speculate that the action localization performance in the ActivityNet1.2 dataset depends more on the classification module, which might be the bottleneck for our model. This speculation also correlates with the different λ values in Eq. 9 when calculating localization score on THUMOS14 and ActivityNet1.2 datasets. According to our model’s assumption, key instance assignment score Q_t implies the action clips and higher weight for this part facilitates the localization. On THUMOS14, the weight λ for the key instance assignment score Q_t is set to be a high value 0.8. But for ActivityNet1.2, the classification score $P_{t,c}$ has a higher weight (0.7), implying that the model mostly relies on classification to succeed on this dataset. For further illustration, we also visualize some good and bad detection results from ActivityNet1.2 dataset in Fig. 5(b).

4.3 Ablation Studies

We ablate our pseudo label generation scheme and Expectation-Maximization alternating training method on THUMOS14 dataset with mAP@0.5 in Table 3.

Ablation on the Pseudo Labeling: We first ablate on the pseudo labeling scheme for \hat{z}_t and $\hat{y}_{t,c}$, and include the results in Table 3. We switch our learning to be supervised by an attention-MIL loss based on softmax function, similar to [17, 26]. In the E step, classification scores of all classes contribute collectively to the attention weights. In the M step, attention weights are applied equally to both positive and negative videos without paying special attention to the bag’s

Table 3: Ablation results for the pseudo labeling and EM alternating training on THUMOS14 dataset in terms of mAP@0.5 (%).

Ablation Models	Pseudo Label	Alternating Training	mAP@0.5
Alternating model		✓	24.5
Pseudo labeling model	✓		26.8
Full Model	✓	✓	30.5

label. Compared to the “Alternating model” doing alternating training but with a plain attention, “Full Model” improves mAP@0.5 from 24.5% to 30.5%. This indicates the usefulness of the proposed pseudo labeling strategy. It models the key instance assignment explicitly and aligns with the MIL assumption better.

Ablation on the EM Alternating Training Technique: We also evaluate the effectiveness of Expectation-Maximization alternating training compared to joint optimization. The EM training method iteratively estimates the key instance assignment, then maximizes the video classification accuracy, and achieves better activity detection performance. “Full Model” improves mAP@0.5 from 26.8% to 30.5% compared to “Pseudo labeling” model with joint optimization. The same training process can be potentially applied on other MIL based models for weakly-supervised object detection task to improve accuracy as well.

5 Conclusion

We propose a EM-MIL framework with pseudo labeling and alternating training for weakly-supervised action detection in video. Our EM-MIL framework is motivated by traditional MIL literature which is under-explored in deep learning settings. By allowing us to explicitly model latent variables, this framework improves our control over the learning objective of the instance-level MIL, which leads to state of the art performance. While this work uses a relatively simple pseudo-labeling scheme to implement the EM method, more sophisticated EM methods can be designed, e.g. explicitly parameterize the latent distribution for instances and directly optimize the instance likelihood in E and M steps. Incorporating the video’s temporal structure is also a promising direction for further performance improvement.

Acknowledgement

Prof. Darrells group was supported in part by DoD, BAIR and BDD.

References

1. Alwassel, H., Heilbron, F.C., Thabet, A., Ghanem, B.: Refineloc: Iterative refinement for weakly-supervised action localization. arXiv preprint arXiv:1904.00227 (2019) [4](#), [11](#)
2. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognition **77**, 329 – 353 (2018) [9](#)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4724–4733 (2017) [10](#)
4. Dietterich, T., Lathrop, R., Lozano-Perez, T.: Solving the multiple instance problem with axis-parallel rectangles. In: Artificial Intelligence. vol. 89, pp. 31–71 (1997) [1](#), [2](#), [4](#)
5. Dooly, D.R., Zhang, Q., Goldman, S.A., Amar, R.A., Brodley, E., Danyluk, A.: Multiple-instance learning of real-valued data. In: Journal of Machine Learning Research. pp. 3–10. Morgan Kaufmann (2001) [2](#), [5](#)
6. Gao, J., Yang, Z., Nevatia, R.: Cascaded boundary regression for temporal action detection. arXiv preprint arXiv:1705.01180 (2017) [11](#)
7. Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015) [9](#), [10](#)
8. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. arXiv preprint arXiv:1802.04712 (2018) [3](#), [5](#), [8](#)
9. Jiang, Y.G., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: THUMOS Challenge: Action Recognition with a Large Number of Classes. <http://crcv.ucf.edu/THUMOS14/> (2014) [9](#), [10](#)
10. Keeler, J.D., Rumelhart, D.E., Leow, W.K.: Integrated segmentation and recognition of hand-printed numerals. In: Lippmann, R.P., Moody, J.E., Touretzky, D.S. (eds.) Advances in Neural Information Processing Systems 3, pp. 557–563. Morgan-Kaufmann (1991) [5](#)
11. Li, X., Kan, M., Shan, S., Chen, X.: Weakly supervised object detection with segmentation collaboration. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) [2](#)
12. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: Bsn: Boundary sensitive network for temporal action proposal generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018) [11](#)
13. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1298–1307 (2019) [2](#), [3](#)
14. Liu, Z., Wang, L., Zhang, Q., Gao, Z., Niu, Z., Zheng, N., Hua, G.: Weakly supervised temporal action localization through contrast based evaluation networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3899–3908 (2019) [4](#), [11](#), [13](#)
15. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: Jordan, M.I., Kearns, M.J., Solla, S.A. (eds.) Advances in Neural Information Processing Systems 10, pp. 570–576. MIT Press (1998) [4](#)
16. Narayan, S., Cholakkal, H., Khan, F.S., Shao, L.: 3c-net: Category count and center loss for weakly-supervised action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8679–8687 (2019) [11](#), [13](#)

17. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6752–6761 (2018) [2](#), [3](#), [4](#), [5](#), [6](#), [9](#), [11](#), [13](#)
18. Nguyen, P.X., Ramanan, D., Fowlkes, C.C.: Weakly-supervised action localization with background modeling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5502–5511 (2019) [6](#), [11](#)
19. Paul, S., Roy, S., Roy-Chowdhury, A.K.: W-talc: Weakly-supervised temporal activity localization and classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 563–579 (2018) [3](#), [5](#), [9](#), [11](#), [12](#), [13](#)
20. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: 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. pp. 5734–5743 (2017) [11](#)
21. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 154–171 (2018) [4](#), [9](#), [11](#), [12](#), [13](#)
22. Singh, K.K., Lee, Y.J.: 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). pp. 3544–3553. IEEE (2017) [3](#), [11](#)
23. Su, H., Zhao, X., Lin, T.: Cascaded pyramid mining network for weakly supervised temporal action localization. In: Asian Conference on Computer Vision. pp. 558–574. Springer (2018) [3](#)
24. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2843–2851 (2017) [4](#)
25. Wan, F., Liu, C., Ke, W., Ji, X., Jiao, J., Ye, Q.: C-mil: Continuation multiple instance learning for weakly supervised object detection. In: CVPR. pp. 2199–2208 (2019) [2](#)
26. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017) [2](#), [3](#), [5](#), [8](#), [9](#), [11](#), [13](#)
27. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016) [10](#)
28. Xu, H., Das, A., Saenko, K.: R-c3d: Region convolutional 3d network for temporal activity detection. In: Proceedings of the IEEE international conference on computer vision. pp. 5783–5792 (2017) [11](#)
29. Xu, H., Das, A., Saenko, K.: Two-stream region convolutional 3d network for temporal activity detection. IEEE transactions on pattern analysis and machine intelligence **41**(10), 2319–2332 (2019) [11](#)
30. Yu, T., Ren, Z., Li, Y., Yan, E., Xu, N., Yuan, J.: Temporal structure mining for weakly supervised action detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5522–5531 (2019) [11](#), [13](#)
31. Yuan, Y., Lyu, Y., Shen, X., Tsang, I.W., Yeung, D.Y.: Marginalized average attentional network for weakly-supervised learning. arXiv preprint arXiv:1905.08586 (2019) [2](#), [3](#)
32. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Joint pattern recognition symposium. pp. 214–223. Springer (2007) [10](#)

33. Zeng, R., Gan, C., Chen, P., Huang, W., Wu, Q., Tan, M.: Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing* **28**(12), 5797–5808 (2019) [3](#)
34. Zhang, C., Platt, J.C., Viola, P.A.: Multiple instance boosting for object detection. In: Weiss, Y., Schölkopf, B., Platt, J.C. (eds.) *Advances in Neural Information Processing Systems* 18, pp. 1417–1424. MIT Press (2006) [5](#)
35. Zhang, Q., Goldman, S.A.: Em-dd: An improved multiple-instance learning technique. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) *Advances in Neural Information Processing Systems* 14, pp. 1073–1080. MIT Press (2002) [2, 5](#)
36. Zhang, Q., Goldman, S.A.: Em-dd: An improved multiple-instance learning technique. In: *Advances in neural information processing systems*. pp. 1073–1080 (2002) [5](#)
37. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2914–2923 (2017) [11, 13](#)
38. Zhong, J.X., Li, N., Kong, W., Zhang, T., Li, T.H., Li, G.: Step-by-step erasure, one-by-one collection: A weakly supervised temporal action detector. arXiv preprint arXiv:1807.02929 (2018) [3](#)