Dual-Evidential Learning for Weakly-supervised Temporal Action Localization

Mengyuan Chen[1,2], Junyu Gao[1,2], Shicai Yang[3], and Changsheng Xu[1,2,4]

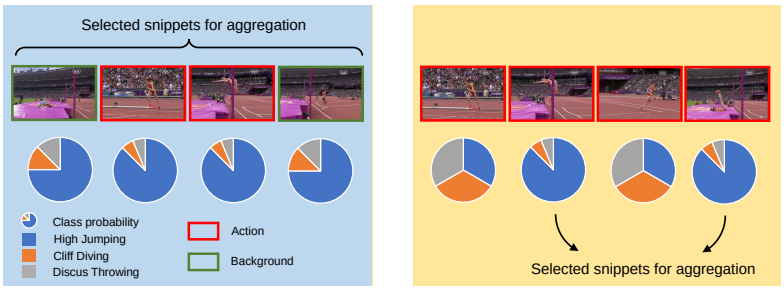
[1] National Lab of Pattern Recognition (NLPR),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Hikvision Research Institute, Hangzhou, China
[4] Peng Cheng Laboratory, ShenZhen, China
chenmengyuan2021@ia.ac.cn; {junyu.gao, csxu}@nlpr.ia.ac.cn;
yangshicai@hikvision.com

Abstract. Weakly-supervised temporal action localization (WS-TAL) aims to localize the action instances and recognize their categories with only video-level labels. Despite great progress, existing methods suffer from severe action-background ambiguity, which mainly comes from background noise introduced by aggregation operations and large intra-action variations caused by the task gap between classification and localization. To address this issue, we propose a generalized evidential deep learning (EDL) framework for WS-TAL, called Dual-Evidential Learning for Uncertainty modeling (DELU), which extends the traditional paradigm of EDL to adapt to the weakly-supervised multi-label classification goal. Specifically, targeting at adaptively excluding the undesirable background snippets, we utilize the video-level uncertainty to measure the interference of background noise to video-level prediction. Then, the snippet-level uncertainty is further deduced for progressive learning, which gradually focuses on the entire action instances in an "easy-to-hard" manner. Extensive experiments show that DELU achieves state-of-the-art performance on THUMOS14 and ActivityNet1.2 benchmarks. Our code is available in github.com/MengyuanChen21/ECCV2022-DELU.

Keywords: Weakly-supervised temporal action localization, Evidential deep learning, Action-background ambiguity

1 Introduction

Temporal action localization is one of the most fundamental tasks of video understanding, which aims to localize the start and end timestamps of action instances and recognize their categories simultaneously in untrimmed videos [62, 49, 31, 53]. It has attracted significant attention from both academia and industry, due to the great potential for video retrieval [9, 41], summarization [24], surveillance [18, 51], anomaly detection [50], visual question answering [25], to name a few. In recent years, numerous action localization methods have been proposed and achieved remarkable performance under the fully-supervised setting.



(a) Background noise (b) Large intra-action variations

Fig. 1: Action-background ambiguity in WS-TAL. (a) Some background snippets are misclassified to the foreground, thus distracting the aggregation process under video-level supervision. (b) Due to the large intra-action variations, the learned action classifiers tend to ignore snippets which are not discriminative enough, thus easily responding to only a fraction of action snippets instead of the entire action instances.

However, these methods require extensive manual frame-level annotations, which limits their scalability and practicability in real-world application scenarios, since densely annotating large amounts of videos is time-consuming, error-prone and extremely costly. To address this problem, weakly-supervised temporal action localization (WS-TAL) methods have been explored [13, 14, 42, 56], which requires only easily available video-level labels.

Due to the absence of frame-wise annotations in the weakly-superwised setting, most existing WS-TAL methods adopt the localization-by-classification strategy [45, 54, 40, 52], in which the commonly used mutiple-instance learning (MIL) strategy [35] and/or attention mechanism [42] are employed. Specifically, after dividing each untrimmed video into multiple fixed-size non-overlapping snippets, these methods apply action classifiers to predict a sequence of classification probabilities of snippets, termed as Class Activation Sequence (CAS). The top ranked snippets are then selected for aggregation, resulting in a video-level prediction for model optimization. To improve the accuracy of the learned CAS, a variety of strategies have been adopted, such as feature enhancement [57, 13], pseudo label generation [56], context modeling [42], contrastive learning [58], which have achieved impressive performance.

Despite remarkable progress has been achieved, existing methods still suffer from severe action-background ambiguity due to the weakly-supervised setting, thus leading to the significant performance gap with fully-supervised methods [26, 27, 29]. We argue that the action-background ambiguity mainly comes from two aspects: **(1)** Background noise introduced by the aggregation operations when generating video-level predictions. As shown in Figure 1(a), the selection of the top action snippets for later aggregation may be inaccurate, *i.e.* some background snippets are mistakenly recognized as action snippets due

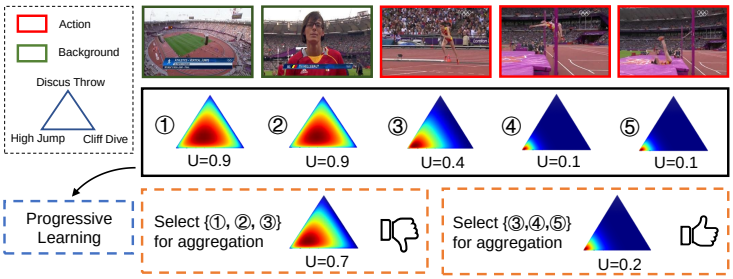


Fig. 2: A toy example of 3-class WS-TAL, which demonstrates the two-level evidential learning structure of DELU. (1) The video-level uncertainty is learned to adaptively exclude the undesirable background snippets in the aggregation process (Section 3.3). (2) The snippet-level uncertainty is employed to better perform foreground-background separation by progressive learning (Section 3.4). Each triangle in this figure represents a Dirichlet distribution of all possible prediction results(Section 3.1). The three vertices of the triangle represent three categories and each point in the triangle represents a particular allocation of class probabilities. When points with high values are concentrated at a certain vertex, the model classifies the sample into the corresponding category with a low uncertainty U.

to their similarity with the foreground in appearance. As a result, background noise will distract or even dominate the further video-level classification. **(2)** Large intra-action variations caused by the task gap between classification and localization. Since only video-level supervisions are provided for WS-TAL, the learned classifiers only need to focus on the most discriminative action snippets when performing video-level classification. As shown in Figure 1(b), the model tends to ignore the action snippets that are not significant enough, *i.e.* fail to classify them into the target action category, thus easily responding to only a fraction of action snippets instead of the entire action instance. These two issues are essentially entangled with each other, jointly intensifying the action-background ambiguity in the model learning process.

Inspired by the above observation, we find that it is desirable to tackle the action-background ambiguity by considering the uncertainty of classification results in both video and snippet levels. Recently, Evidential Deep Learning (EDL) [44, 36], which can collect the evidence of each category and quantify the predictive uncertainty, has received extensive attention and achieved impressive performance in a few computer vision tasks [46, 43, 3]. However, EDL is designed for fully-supervised single-label classification tasks, which is not suitable to be directly integrated into weakly-supervised temporal action localization[5].

To address the above issues, we propose a generalized EDL framework for WS-TAL, called Dual-Evidential Learning for Uncertainty modeling (DELU),

[5] In WS-TAL, multiple types of action may appear simultaneously in a video.

which extends the traditional paradigm of EDL to adapt to the weakly-supervised multi-label classification goal. As shown in Figure 2, to tackle the action-background ambiguity, DELU leverages a two-level evidential learning structure to model the predictive uncertainty in both video level and snippet level. Specifically, **(1)** we utilize the video-level uncertainty to measure the interference of background noise. Here, we propose a novel evidential learning objective to learn the video-level unceratinty, which can adaptively exclude the undesirable background snippets in the aggregation operations. **(2)** When pursuing video-level uncertainty, the snippet-level uncertainty is naturally deduced. Based on this more fine-grained information, we design a progressive learning strategy, in which the order of the snippet-level uncertainty is leveraged to gradually focus on the entire action instances in an "easy-to-hard" manner. As a result, the negative impact of intra-action variations is alleviated and the background noise can be further excluded. Our proposed DELU is optimized in an end-to-end manner, and we validate its effectiveness on two popular benchmarks [15, 6].

In conclusion, the main contributions of this work are three-fold:

1. We design a generalized EDL paradigm to better adapt to the multi-label classification setting under weak supervision. To the best of our knowledge, we are among the first to introduce the evidential deep learning to weakly-supervised temporal action localization.

2. By carefully considering both video- and snippet-level uncertainty, we propose a novel dual-evidential learning framework, which can effectively alleviates the action-background ambiguity caused by background noise and large intra-action variations.

3. We conduct extensive experiments on two public benchmarks, *i.e.*, Thumos14 dataset and Activity1.2 dataset. On both benchmarks our proposed DELU method achieves state-of-the-art results.

2 Related Work

Weakly-supervised Temporal Action Localization (WS-TAL). In recent years, WS-TAL with various types of weak supervisions has been developed, *e.g.*, action orders [5], web videos [11], single-frame annotation [34, 21], and video-level action category labels [52, 39, 28], while the last one is the most commonly adopted due to its simplicity. UntrimmedNet [52] is the first work to use video-level action category labels for the WS-TAL task. To date in the literature, most existing approaches can be divided into three categories, namely attention-based methods [45, 54, 13, 42, 38, 32], MIL-based methods [35, 22, 33, 37, 40], and erasing-based methods [48, 59, 61]. Attention-based approaches generate the foreground attention weight to suppress the background parts. CO2-Net [13] filters out the information redundancy to enhance features by cross-modal attention alignment. MIL-based approaches treat the input video as a bag in which the action clips are positive samples and the background clips are negative ones, and a top-k operation is utilized to aggregate the snippet-level prediction results. ASL [35] explores a general independent concept of action by investigating a

class-agnostic actionness network. The erasing-based methods attempt to erase the most discriminative parts to highlight other less discriminative snippets. For example, FC-CRF [61] tries to find new foreground snippets progressively via step-by-step erasion from a complete input video.

Although several methods have investigated the role of uncertainty in WS-TAL, *e.g.*, GUCT [56] estimates the uncertainty about the generated snippet-level pseudo labels to mitigate noise, Lee *et al.* [23] decompose the classification probability into the action probability and the uncertainty with a chain rule, they neglect the unique two-level uncertainty structure under the weakly-supervised setting of WS-TAL. In this paper, by carefully considering both video- and snippet-level uncertainty, we propose a novel dual-evidential learning framework to effectively alleviate the action-background ambiguity.

Evidential Deep Learning (EDL). In recent years, deep learning approaches commonly adopt softmax function as the classification head to output final predictions. However, due to the exponent operation employed on neural network outputs, there exist intrinsic deficiencies of modeling class probabilities with softmax. On the one hand, softmax-based classifiers have a tendency to be over-confident in false predictions, which brings additional difficulties to the optimization process [12]. On the other hand, since the softmax output is essentially a point estimate of the probability distribution [10], it cannot estimate the predictive uncertainty.

To overcome the above weaknesses, EDL [44, 36] was gradually developed and refined based on Dempster-Shafer theory of evidence (DST) [55] and Subjective Logic theory [19]. The core idea of EDL is to collect evidence of each category and construct a Dirichlet distribution parametrized over the collected evidence to model the distribution of class probabilities. Besides the probability of each category, the predictive uncertainty can be quantified from the distribution by Subjective Logic theory. EDL has been successfully utilized in various tasks requiring uncertainty modeling, and remarkable progress has been achieved in a few computer vision tasks [46, 43, 3]. For example, Bao *et al.* [3] use the uncertainty obtained by EDL to distinguish between the known and unknown samples for the open set action recognition (OSAR) task.

However, current EDL models are designed for fully-supervised single-label classification tasks, which is not suitable to be directly integrated into weakly-supervised multi-label classification setting. To the best of our knowledge, we are among the first to introduce the evidential deep learning to the WS-TAL task, demonstrating favorable performance.

3 Proposed Approach

In this work, we describe our DELU framework in details. We first introduce the Evidential Deep Learning (EDL) in Section 3.1. The overview architecture of DELU is illustrated in Figure 3. We firstly utilize a pre-trained feature extractor to obtain snippet-level video feature and adopt a backbone network to obtain the CAS (Section 3.2). Then, we propose a generalized EDL paradigm which can

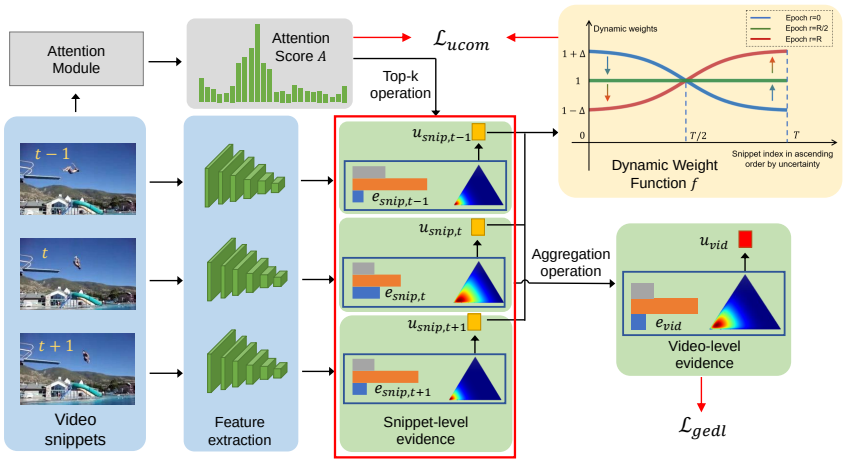


Fig. 3: Overall framework of the proposed DELU. After obtaining the snippet-level evidence, we aggregate them to generate the video-level evidence by selecting the top-k snippets according to the attention scores. The video-level evidence and uncertainty are used to generalize the EDL paradigm for WS-TAL, and the snippet-level uncertainty is employed to generate dynamic weights for progressive learning. Note that we omit the regular classification loss L_{cls} (Section 3.2) in this figure for simplicity.

better adapt to the setting of WS-TAL. Specifically, the video-level uncertainty is utilized to generalize the EDL paradigm for weakly-supervised multi-label (WS-Multi) classification (Section 3.3), and a progressive learning strategy is employed by leveraging the snippet-level uncertainty (Section 3.4). Finally, the whole framework is end-to-end learned (Section 3.5).

3.1 Background of Evidential Deep Learning

According to Dempster-Shafer theory of evidence [55] and Subjective Logic theory [19], evidential deep learning (EDL) [2, 44] was proposed to address the deficiencies of softmax-based classifiers mentioned in Section 2. Instead of directly predicting the probability of each class, EDL collects evidence of each class first and then builds a Dirichlet distribution of class probabilities parametrized over the collected evidence. Based on the distribution, the predictive uncertainty can be quantified by Subjective Logic theory [19]. To represents the intensity of activation of each class, *evidence* is defined as a measure of the amount of support collected from data in favor of a sample being classified into a particular class [55, 19, 44].

EDL targets at predicting evidence for each category and building a Dirichlet distribution of class probability. Given a C-class classification problem, let $\boldsymbol{e} \in \mathbb{R}_{+}^{C}$ be the evidence vector predicted for a sample x, the corresponding Dirichlet

distribution is given by

$$D(\boldsymbol{q}\boldsymbol{j}\ \ ) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{C} q_j^{\alpha_j\ \ 1}, & \text{for } \boldsymbol{q}\ \boldsymbol{2}\ \boldsymbol{S}_C \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $\alpha_j = e_j + 1$, $j = 1, ..., C$ is the class index, B denotes the C-dimensional beta function and $\boldsymbol{q}$ is a point on the C-dimensional unit simplex $\boldsymbol{S}_C$ [44]. As shown in Figure 2 and Figure 3, the Dirichlet distribution over a three-dimensional simplex can be visualized as a triangle heatmap. Eac h point of the simplex represents a point estimate of the probability distribution, and each edge is the value range $[0, 1]$, while the brightness represents the value of the Dirichlet probability density function. Treating $D(\boldsymbol{q}\boldsymbol{j}\ \ )$ as the class probability distribution, the negative logarithm of the marginal likelihood for sample x can be derived as follows:

$$\boldsymbol{L}_{EDL} = \sum_{j=1}^{C} y_j(\log S \quad \log \alpha_j), \tag{2}$$

where $\boldsymbol{y}$ is the one-hot ground-truth vector for sample x, $S = \sum_{j=1}^{C} \alpha_j$. Eq. (2) is the traditional optimization objective of EDL [44, 36]. Then, the predicted probablity $\hat{p}_j$ of class j and the uncertainty u of the prediction can be derived as following:

$$\hat{p}_j = \alpha_j / S, \quad u = C/S. \tag{3}$$

Note that uncertainty u is inversely proportional to the total evidence. When the total evidence is zero, the uncertainty becomes the maximum.

3.2 Notations and Preliminaries

In the following, superscript (i) is used to indicate the sample index, $i = 1, ..., N$, and subscript j is used to indicate the category index. Note that in the following, for simplicity, the superscript (i) has been omitted when there is no ambiguity. Given an untrimmed video V and its corresponding multi-hot action category label $\boldsymbol{y}\ \boldsymbol{2}\ \boldsymbol{f}0, 1\boldsymbol{g}^{C+1}$, where C is the action category number, and $C + 1$ represents the non-action background class. The action instances in video V detected by WS-TAL methods can be formulated as a set of ordered quadruplets $\boldsymbol{f}c_m, t_m^s, t_m^e, \phi_m\boldsymbol{g}_{m=1}^{M}$, where M is the number of action instances in V, c_m denotes the action category, t_m^s and t_m^e denote the start and end timestamps, and ϕ_m denotes the confidence score.

Following previous works [42, 14, 56], we first divide the untrimmed video V into T non-overlapping 16-frame snippets, and use pre-trained networks, *e.g.*, I3D model [20], to extract features from both RGB and optical flow streams. After that, the two types of features are concatenated and then fed into an fusion module, *e.g.*, convolutional layers [42, 13], to obtain the snippet-wise feature $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_T]\ \boldsymbol{2}\ \boldsymbol{R}^{D\ \ T}$, where D is the feature dimension.

To date in the literature, existing methods mainly embrace a localization-by-classification strategy. Firstly, a classifier f_{cls} is applied to the snippet-wise features $\boldsymbol{X}$ to predict the CAS, denoted as $\boldsymbol{p} = [\boldsymbol{p}_1, ..., \boldsymbol{p}_T] \in \mathbb{R}^{T \times (C+1)}$. Meanwhile, an attention score sequence $\boldsymbol{A} = [A_1, ..., A_T] \in \mathbb{R}^T$ is predicted by an attention module to represent the probabilities of snippets belonging to the foreground. After that, the video-level classification probability $\tilde{\boldsymbol{y}}$ is obtained through a top-k aggregation operation over the CAS according to the attention scores $\boldsymbol{A}$, and the process can be formalized as:

$$\tilde{\boldsymbol{y}} = \frac{1}{k} \sum_{\substack{t \in \Omega, |\Omega| = k, \\ \Omega = \arg\max_{\Omega} \sum_{t \in \Omega} A_t}} \boldsymbol{p}_t, \tag{4}$$

where $\boldsymbol{p} = f_{cls}(\boldsymbol{X})$. Finally, the video-level prediction $\tilde{\boldsymbol{y}}$ is optimized by the ground-truth label $\boldsymbol{y}$:

$$\boldsymbol{L}_{cls} = \text{Cross-entropy}(\boldsymbol{y}, \tilde{\boldsymbol{y}}). \tag{5}$$

3.3 Generalizing EDL for Video-level WS-Multi Classification

Although evidential deep learning has made great progress in modeling uncertainty, the traditional EDL paradigm is not suitable to be directly applied to the WS-Multi classification setting of WS-TAL. In order to extend the applicability of evidential learning methods to WS-TAL tasks, the first problem to be solved is how to predict video-level evidence $\boldsymbol{e}_{vid} = [\boldsymbol{e}_{vid,1}, ..., \boldsymbol{e}_{vid,C}] \in \mathbb{R}^C$ from snippet-level features X. We propose to predict the snippet-level evidence $\boldsymbol{e}_{snip} = [\boldsymbol{e}_{snip,1}, ..., \boldsymbol{e}_{snip,C}] \in \mathbb{R}^{T \times C}$ for action categories first, and then obtain the video-level evidence $\boldsymbol{e}_{vid}$ by aggregating the snippet-level evidence $\boldsymbol{e}_{snip,t}$ of the snippets which are attached with the top-k attention score A_t. Note that here we jointly employ attention scores and evidence for aggregation, which makes the attention module and evidence learning enhance and complement each other. Formally, we can denote the video-level evidence collection process as following:

$$\boldsymbol{e}_{vid} = \frac{1}{k} \sum_{\substack{t \in \Omega, |\Omega| = k, \\ \Omega = \arg\max_{\Omega} \sum_{t \in \Omega} A_t}} \boldsymbol{e}_{snip,t}, \tag{6}$$

where $k = \lceil T/r \rceil$, r is a scaling factor, $\boldsymbol{e}_{snip} = g(f(\boldsymbol{X}; \quad))$, f is a DNN parameterized by $\quad$ to collect evidence, g denotes an evidence function, *e.g.*, ReLU, to keep the evidence $\boldsymbol{e}_{snip}$ non-negative. Note that here we only consider the C action categories for evidential learning since the additional background class hinders the uncertainty modeling of foreground. Following the traditional EDL method, we obtain α, S, and u_{vid} by

$$\alpha_j = e_{vid,j} + 1, S = \sum_{j=1}^{C} \alpha_j, u_{vid} = C/S, \tag{7}$$

Due to the low frequency or short duration of some action categories, the collected evidence for them tend to have a relatively low intensity, thus being easily ignored in the process of model learning. Therefore, we hope that the classifier can assign more importance to the target action categories with smaller evidence scores. With the symbols introduced in Section 3.2, we design a new label vector $\boldsymbol{g}$ to replace original multi-hot label $\boldsymbol{y}$:

$$g_j = \frac{y_j / e_{vid,j}}{\sum_{j=1}^{C} y_j / e_{vid,j}}, \tag{8}$$

It can be found from Eq. (8) that g_j and $e_{vid,j}$ are inversely proportional, thus the model can learn features of each target category more evenly.

Although the above modified EDL paradigm can better adapt to the multi-label classification setting, it neglects the uncertainty derived from the video-level evidential learning. We further notice that the video-level uncertainty $\boldsymbol{u}_{vid}$ can be utilized to measure the interference of background noise to video-level prediction, thus avoiding the background noise intensifying the action-background ambiguity. We argue that the selected top-k snippets are dominated by action snippets as expected only when the classifier predicts the video category correctly with a low uncertainty. Contrarily, when the prediction is accompanied by a high uncertainty, the video-level prediction is more likely to be dominated by background noise. In the latter case, we should expect the classifier to produce a trivial prediction, instead of forcing the result to be consistent with the given video-level action category label, which may lead to the action-background ambiguity further increasing. To achieve this goal, we propose to replace g_j with h_j by leveraging the video level uncertainty:

$$h_j = (1 \quad u_{vid}) g_j, \tag{9}$$

Therefore, the samples with higher video-level uncertainties can take a smaller weight in the optimization process, thus reducing the negtive impact caused by background noise.

Based on the above derivation, our objective for generalizing EDL can be formulated to the following form:

$$\boldsymbol{L}_{gedl} = \sum_{i=1}^{N} (1 \quad u_{vid}^{(i)}) \sum_{j=1}^{C} \frac{y_j^{(i)} / e_j^{(i)}}{\sum_{j=1}^{C} y_j^{(i)} / e_j^{(i)}} (\log S^{(i)} \quad \log \alpha_j^{(i)}). \tag{10}$$

3.4 Snippet-level Progressive Learning

In the above section, snippet-level uncertainty is also deduced when performing video-level evidential learning. To leverage the fine-grained information, we notice that $\boldsymbol{p} \; 2 \; \mathsf{R}^{T \;\; (C+1)}$ represents the classification probabilities of snippets, and $p_{t,c+1}$ indicates the probability of the t-th snippet belonging to the background. It is natural to think that the attention score $\boldsymbol{A}$, which represents the probability of each snippet belonging to the foreground, and the background probability

$p_{t,c+1}$ should be complementary:

$$L_{com} = \sum_{t=1}^{T} jA_t + p_{t,c+1} \quad 1j, \tag{11}$$

where $j \ j$ is the ℓ_1 norm.

Due to the existence of task gap between classification and localization, models tend to focus only on the most discriminative video snippets, which makes it difficult to classify other action snippets correctly. Inspired by Curriculum Learning [4], we propose a progressive learning method by leveraging snippet-level uncertainty to help the model learn the entire action instance progressively and comprehensively. Note that the snippet-level uncertainty can reflect the discriminability of itself, that is, the lower uncertainty of an action snippet means it is easier to recognize its category. Our strategy is to attach larger weights to snippets with lower uncertainty and smaller weights to ones with higher uncertainty in the beginning, and gradually reverse this allocation in the training process. During the progressively learning, the model firstly focuses on easy action snippets and then gradually pays more attention to background and difficult action snippets. As a result, the negtive impact of intra-action variation is alleviated and the background noise can be further excluded. Therefore, as shown in Figure 3, we design a dynamic weight function $\lambda(r,t)$ as following:

$$\lambda(r,t) = \Delta \ \tanh\left(\delta(r)\phi(s(t))\right) + 1, \tag{12}$$

where Δ is a hyper-parameter representing the amplitude of the change of the dynamic weights. Specifically, $\delta(r) = \frac{2r}{R} \quad 1 \ 2 \ [\ 1,1]$, $r = 1,...,R$, r is the current epoch index, R denotes the total training epoch number, and $\phi(s(t)) = \frac{2s(t)}{T} \quad 1 \ 2 \ [\ 1,1]$, $s = 1,...,T$, $s(t)$ indicates the ordinal number of snippet t obtained by sorting the snippet-level uncertainty u_{snip} in a descending order.

Finally, after multiplying the snippet-level uncertainty guided dynamic weights to the complementary loss L_{com}, we can gradually focus on the entire action instances in an "easy-to-hard" manner by optimizing the following objective:

$$L_{ucom} = \sum_{t=1}^{T} \lambda(r,t) \ \ jA_t + p_{c+1,t} \quad 1j. \tag{13}$$

3.5 Learning and Inference

Training. By aggregating all the aforementioned optimization objectives, we obtain the final loss function as following

$$L = L_{cls} + \lambda_1 L_{gedl} + \lambda_2 L_{ucom}, \tag{14}$$

here, λ_1, λ_2 are balancing hyper-parameters.

Inference. In the inference phase, we first predict the CAS of the test video and then apply a threshold strategy to obtain action snippet candidates following the standard process [13]. Finally, continuous snippets are grouped into action proposals, and then non-maximum-suppression (NMS) is performed to remove duplicated proposals.

Table 1: Temporal action localization performance comparison with existing methods on the THUMOS14 dataset.

Supervision	Method	mAP@t-IoU(%)									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	[0.1:0.5]	[0.3:0.7]	Avg
Fully	TAL-Net[8], CVPR2018	59.8	57.1	53.2	48.5	42.8	33.8	20.8	52.3	39.8	45.1
	GTAN[31], CVPR2019	69.1	63.7	57.8	47.2	38.8	-	-	55.3	-	-
	BU-TAL[60], ECCV2020	-	-	53.9	50.7	45.4	38.0	28.5	-	43.3	-
Weakly	UntrimmedNet[52], CVPR2017	44.4	37.7	28.2	21.1	13.7	-	-	29.0	-	-
	Hide-and-Seek[48], ICCV2017	36.4	27.8	19.5	12.7	6.8	-	-	20.6	-	-
	AutoLoc[47], ECCV2018	-	-	35.8	29.0	21.2	13.4	5.8	-	-	-
	STPN[39], CVPR2018	52.0	44.7	35.5	25.8	16.9	9.9	4.3	35.0	18.5	27.0
	W-TALC[40], ECCV2018	55.2	49.6	40.1	31.1	22.8	-	7.6	39.8	25.4	34.4
	DGAM[45], CVPR2020	60.0	54.2	46.8	38.2	28.8	19.8	11.4	45.6	29.0	37.0
	RefineLoc[1], WACV2021	-	-	40.8	32.7	23.1	13.3	5.3	-	23.0	-
	ACSNet[30], AAAI2021	-	-	51.4	42.7	32.4	22.0	11.7	-	32.0	-
	HAM-Net[16], AAAI2021	65.9	59.6	52.2	43.1	32.6	21.9	12.5	50.7	32.5	41.1
	ASL[35], CVPR2021	67.0	-	51.8	-	31.1	-	11.4	-	-	40.3
	CoLA[58], CVPR2021	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
	AUMN[32], CVPR2021	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	32.4	41.5
	UGCT[56], CVPR2021	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6	43.6
	D2-Net[38], ICCV2021	65.7	60.2	52.3	43.4	36.0	-	-	51.5	-	-
	FAC-Net[14], ICCV2021	67.6	62.1	52.6	44.3	33.4	22.5	12.7	52.0	33.1	42.2
	ACM-Net[42], arXiv2021	68.9	62.7	55.0	44.6	34.6	21.8	10.8	53.2	33.4	42.6
	CO2-Net[13], MM2021	70.1	63.6	54.5	45.7	38.3	26.4	13.4	54.4	35.7	44.6
	ACG-Net[57], AAAI2022	68.1	62.6	53.1	44.6	34.7	22.6	12.0	52.6	33.4	42.5
	DELU(Ours)	**71.5**	**66.2**	**56.5**	**47.7**	**40.5**	**27.2**	**15.3**	**56.5**	**37.4**	**46.4**

4 Experimental Results

We evaluate our proposed DELU on two public benchmarks, *i.e.*, THUMOS14 [15] and ActivityNet1.2 [6]. The following experiments verifies the effectiveness.

4.1 Experimental Setup

THUMOS14. It contains 200 validation videos and 213 test videos annotated with temporal action boundaries from 20 action categories. Each video contains an average of 15.4 action instances, making this dataset challenging for weakly-supervised temporal action localization.

ActivityNet1.2. ActivityNet1.2 contains 4,819 training and 2,383 validation videos from 100 action categories. Since the ground-truth annotations of the test set is not yet public, we test on the validation set following the protocol in previous work [17, 16, 13].

Evaluation Metrics. Following previous work [13, 42, 52], we use mean Average Precision (mAP) under different temporal Intersection over Union (t-IoU) thresholds as evaluation metrics. The t-IoU thresholds for THUMOS14 is [0.1:0.1:0.7] and for ActivityNet is [0.5:0.05:0.95].

Implementation Details. Following existing methods, we use I3D [7] model pretrained on Kinetics [20] dataset to extract both the RGB and optical flow

features. After that, we adopt CO2-Net [13] as the backbone to obtain the fused 2048 dimensional features and implement f_{cls} and $\mathcal{L}_{cls}$. The number of the sampled snippets T for THUMOS14 and ActivityNet1.2 is set to 320 and 60, and the scaling factor r is set to 7 and 5, respectively. Two convolutional layers are utilized as the evidence collector f. The amplitude Δ is set to 0.7, and the balancing hyper-parameters λ_1 and λ_2 are 1.3 and 0.4.

4.2 Comparision with State-of-the-Art Methods

Evaluation on THUMOS14. Table 1 compares DELU with existing fully and weakly-supervised TAL methods on the THUMOS14 dataset. From this table we can find that DELU outperforms all existing weakly-supervised methods in all IoU metrics. Specifically, our method achieves impressive performance of 15.3% mAP@0.7 and 46.4% mAP@Avg, and an absolute gain of 1.8% and 2.8% is obtained in terms of the average mAP when compared to the SOTA approaches CO2-Net [13] and UGCT [56]. In addition to this, we observe that our methods can even achieve comparable performance with those fully-supervised methods, especially in terms of metrics with low IoU.

Evaluation on ActivityNet1.2. Table 2 presents the comparison of experimental performance on the ActivityNet1.2 dataset. As shown, our method also achieves state-of-the-art performance under the weakly-supervised setting. Specifically, compared with the state-of-the-art method ACM-Net[42], we obtain a relative gain of 1.5% in the term of the average mAP. DELU achieves less significant performance improvement on this dataset due to the different characteristics of datasets, that THUMOS14 contains 15.4 action instances per video on average, compared with 1.6 in each video of ActivityNet. Therefore, methods that tend to treat ambiguous snippets as the foreground will perform better on ActivityNet, while methods with the opposite tendency will achieve better performance on THUMOS14. For example, ACM-Net achieves SOTA on ActivityNet, and CO2-Net achieves SOTA on THUMOS14, but neither of them can achieve the same outstanding results on the other dataset. In this paper, the proposed DELU achieves the SOTA performance on both datasets consistently.

4.3 Ablation Study

In Table 3, we investigate the contribution of each component on the THUMOS14 dataset. As introduced in Section 3.5, the optimization objective of our proposed DELU consists of three loss functions, *i.e.*, $\mathcal{L}_{cls}$, $\mathcal{L}_{gedl}$ and $\mathcal{L}_{ucom}$. Firstly, we set the baseline of the ablation study as the backbone method CO2-net [13] whose optimization objective is $\mathcal{L}_{cls}$. On this basis, we conduct experiments on each improvement scheme according to the derivation steps of $\mathcal{L}_{gedl}$, that is, (T) only using the traditional EDL method optimized by Eq. (2), (B) applying the balanced improvement given by Eq. (8) on the basis of the traditional EDL, and (U) optimizing the complete $\mathcal{L}_{gedl}$ which considers the video-level uncertainty (Eq. (10)). Finally, the snippet-level uncertainty guided complementary

Table 2: Comparison results on the ActivityNet1.2 dataset.

Method	mAP@t-IoU(%)			
	0.5	0.75	0.95	Avg
DGAM[45], CVPR2020	41.0	23.5	5.3	24.4
RefineLoc[1], WACV2020	38.7	22.6	5.5	23.2
ACSNet[30], AAAI2021	40.1	26.1	6.8	26.0
HAM-Net[16], AAAI2021	41.0	24.8	5.3	25.1
Lee et al[21], AAAI2021	41.2	25.6	6.0	25.9
ASL[35], CVPR2021	40.2	-	-	25.8
CoLA[58], CVPR2021	42.7	25.7	5.8	26.1
AUMN[32], CVPR2021	42.0	25.0	5.6	25.5
UGCT[56], CVPR2021	41.8	25.3	5.9	25.8
D2-Net[38], ICCV2021	42.3	25.5	5.8	26.0
ACM-Net[42], arXiv2021	43.0	25.8	**6.4**	26.5
CO2-Net[13], MM2021	43.3	26.3	5.2	26.4
ACGNet[57], AAAI2022	41.8	26.0	5.9	26.1
DELU(Ours)	**44.2**	**26.7**	5.4	**26.9**

Table 3: Ablation study of the effectiveness of our proposed EDLU on the THUMOS14 dataset. T represents the traditional EDL method (Eq. (2)). B represents the modified EDL (Eq. (8)) which balances the evidence collected for each target category. U is the generalized EDL paradigm leveraging video-level uncertainty (Eq. (10)).

Exp	$\mathcal{L}_{gedl}$			$\mathcal{L}_{ucom}$	mAP@IoU(%)				
	T	B	U		0.1	0.3	0.5	0.7	mAP
1	✗	✗	✗	✗	69.7	54.7	38.2	13.2	44.5
2	✓	✗	✗	✗	68.9	54.8	39.0	14.9	44.9
3	✓	✓	✗	✗	70.4	55.4	38.9	14.7	45.2
4	✓	✓	✓	✗	70.6	56.2	39.2	14.6	45.6
5	✓	✓	✓	✓	**71.5**	**56.5**	**40.5**	**15.3**	**46.4**

loss $\mathcal{L}_{ucom}$ is added to above components. Table 3 clearly demonstrates that every step of our method brings considerable performance improvement on the THUMOS14 dataset.

4.4 Evaluation for insights

In this part, we provide experiment results to illustrate that **(1)** background noises and **(2)** ignoring non-salient action snippets are existing issues in WS-TAL and DELU effectively alleviates them.

The **issue (1)** does exist. Given the test videos of the THUMOS14 dataset, we select the 5% snippets with the highest target CAS, *i.e.*, the Class Activation Score of the Target (ground-truth) action category, predicted and averaged by the SOTA method CO2-Net, of which 22.54% are background snippets (this number drops to 20.92% in our DELU). This fact indicates that existing methods suffer from the background noise issue, which hinders further improvement of localization performance. To verify that DELU alleviates this issue, we utilize Area Under the Receiver Operating Characteristic (AUROC) to evaluate the action-background separation performance. Specifically, we sort snippets according to the learned attention scores and then divide them into action and background with different thresholds. Figure 4(a) presents the ROC curves of DELU and CO2-Net, which shows DELU achieves more accurate action-background separation. To be more precise, DELU has an AUROC of 85.92%, while CO2-Net has an AUROC of 83.96%.

The **issue (2)** also exists. First, we observe the following phenomenon: the target CAS predicted for action snippets in the same video are not evenly distributed. For instance, as shown in Figure 4(b), when CO2-Net works on the

THUMOS14 dataset, 30% snippets with higher target CAS accounts for more than half of the total target CAS. In order to validate that non-salient snippets can bring performance improvement, we divide snippets in each test video into ten intervals according to the target CAS, and then manually correct the target CAS (before softmax normalization) of the snippets in each interval with a fixed amplitude (5 in our experiment). Figure 4(c) shows that the performance improvements in non-salient intervals are generally higher than those in salient ones. Note that using other amplitude values also has similar results. To verify that DELU improves this issue, we compare the target CAS distribution of DELU and CO2-Net. As shown in Figure 4(d), the target CAS distribution of DELU is more uniform, indicating DELU does make the model more comprehensively focus on both salient and non-salient snippets and improve performance. Similar conclusions can also be found in Figure 4(b).

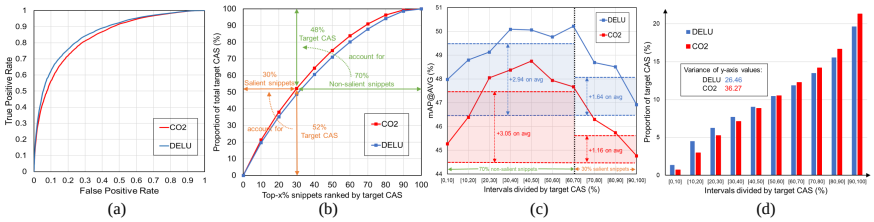


Fig. 4: Evaluation for the insights of our method.

5 Conclusions

This paper proposes a generalized evidential learning framework for WS-TAL, called Dual-Evidential Learning for Uncertainty modeling. Specifically, video-level evidential learning and snippet-level progressive learning are performed to jointly alleviate the action-background ambiguity. Extensive experiments demonstrate the effectiveness of components in our proposed framework. DELU outperforms all existing methods on THUMOS14 and ActivityNet1.2 for weakly-supervised temporal action localization. Inspired by the merits of evidential learning, in the future, we plan to perform pseudo label mining or introduce single-frame annotations, to explore and widen the potential of our DELU framework.

Acknowledgements. This work was supported by the National Key Research & Development Plan of China under Grant 2020AAA0106200, in part by the National Natural Science Foundation of China under Grants 62036012, U21B2044, 61721004, 62102415, 62072286, 61720106006, 61832002, 62072455, 62002355, and U1836220, in part by Beijing Natural Science Foundation (L201001), in part by Open Research Projects of Zhejiang Lab (NO.2022RC0AB02), and in part by CCF-Hikvision Open Fund (20210004).

References

1. Alwassel, H., Heilbron, F.C., Thabet, A., Ghanem, B.: Refineloc: Iterative refinement for weakly-supervised action localization. In: WACV (2019)
2. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. NeurIPS (2020)
3. Bao, W., Yu, Q., Kong, Y.: Evidential deep learning for open set action recognition. In: ICCV (2021)
4. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009)
5. Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. In: ECCV (2014)
6. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
8. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster r-cnn architecture for temporal action localization. In: CVPR (2018)
9. Ciptadi, A., Goodwin, M.S., Rehg, J.M.: Movement pattern histogram for action recognition and retrieval. In: ECCV. Springer (2014)
10. Gal, Y., et al.: Uncertainty in deep learning. PhD thesis, University of Cambridge (2016)
11. Gan, C., Sun, C., Duan, L., Gong, B.: Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In: ECCV (2016)
12. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. PMLR (2017)
13. Hong, F.T., Feng, J.C., Xu, D., Shan, Y., Zheng, W.S.: Cross-modal consensus network for weakly supervised temporal action localization. In: ACM MM (2021)
14. Huang, L., Wang, L., Li, H.: Foreground-action consistency network for weakly supervised temporal action localization. In: ICCV (2021)
15. Idrees, H., Zamir, A.R., Jiang, Y.G., Gorban, A., Laptev, I., Sukthankar, R., Shah, M.: The thumos challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding **155**, 1–23 (2017)
16. Islam, A., Long, C., Radke, R.: A hybrid attention mechanism for weakly-supervised temporal action localization. In: AAAI (2021)
17. Islam, A., Radke, R.: Weakly supervised temporal action localization using deep metric learning. In: WACV (2020)
18. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 221–231 (2012)
19. Jsang, A.: Subjective logic: A formalism for reasoning under uncertainty. Springer Verlag (2016)
20. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv:1705.06950 (2017)
21. Lee, P., Byun, H.: Learning action completeness from points for weakly-supervised temporal action localization. In: ICCV (2021)

22. Lee, P., Uh, Y., Byun, H.: Background suppression network for weakly-supervised temporal action localization. In: AAAI (2020)
23. Lee, P., Wang, J., Lu, Y., Byun, H.: Weakly-supervised temporal action localization by uncertainty modeling. In: AAAI (2021)
24. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 1346–1353. IEEE (2012)
25. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)
26. Li, Z., Yao, L.: Three birds with one stone: Multi-task temporal action detection via recycling temporal annotations. In: CVPR (2021)
27. Lin, C., Xu, C., Luo, D., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., Fu, Y.: Learning salient boundary feature for anchor-free temporal action localization. In: CVPR (2021)
28. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: CVPR (2019)
29. Liu, X., Hu, Y., Bai, S., Ding, F., Bai, X., Torr, P.H.: Multi-shot temporal event localization: a benchmark. In: CVPR (2021)
30. Liu, Z., Wang, L., Zhang, Q., Tang, W., Yuan, J., Zheng, N., Hua, G.: Acsnet: Action-context separation network for weakly supervised temporal action localization. arXiv:2103.15088 (2021)
31. Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Gaussian temporal awareness networks for action localization. In: CVPR (2019)
32. Luo, W., Zhang, T., Yang, W., Liu, J., Mei, T., Wu, F., Zhang, Y.: Action unit memory network for weakly supervised temporal action localization. In: CVPR (2021)
33. Luo, Z., Guillory, D., Shi, B., Ke, W., Wan, F., Darrell, T., Xu, H.: Weakly-supervised action localization with expectation-maximization multi-instance learning. In: ECCV (2020)
34. Ma, F., Zhu, L., Yang, Y., Zha, S., Kundu, G., Feiszli, M., Shou, Z.: Sf-net: Single-frame supervision for temporal action localization. In: ECCV (2020)
35. Ma, J., Gorti, S.K., Volkovs, M., Yu, G.: Weakly supervised action selection learning in video. In: CVPR (2021)
36. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. NeurIPS (2018)
37. Moniruzzaman, M., Yin, Z., He, Z., Qin, R., Leu, M.C.: Action completeness modeling with background aware networks for weakly-supervised temporal action localization. In: ACM MM (2020)
38. Narayan, S., Cholakkal, H., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: D2-net: Weakly-supervised action localization via discriminative embeddings and denoised activations. In: ICCV (2021)
39. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: CVPR (2018)
40. Paul, S., Roy, S., Roy-Chowdhury, A.K.: W-talc: Weakly-supervised temporal activity localization and classification. In: ECCV (2018)
41. Ramezani, M., Yaghmaee, F.: A review on human action analysis in videos for retrieval applications. Artificial Intelligence Review **46**(4), 485–514 (2016)
42. Sanqing Qu, Guang Chen, Z.L.L.Z.F.L.A.K.: Acm-net: Action context modeling network for weakly-supervised temporal action localization. arXiv:2104.02967 (2021)

43. Sensoy, M., Kaplan, L., Cerutti, F., Saleki, M.: Uncertainty-aware deep classifiers using generative models. In: AAAI (2020)
44. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. NeurIPS (2018)
45. Shi, B., Dai, Q., Mu, Y., Wang, J.: Weakly-supervised action localization by generative attention modeling. In: CVPR (2020)
46. Shi, W., Zhao, X., Chen, F., Yu, Q.: Multifaceted uncertainty estimation for label-efficient deep learning. NeurIPS (2020)
47. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: ECCV (2018)
48. Singh, K.K., Lee, Y.J.: Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In: ICCV (2017)
49. Sridhar, D., Quader, N., Muralidharan, S., Li, Y., Dai, P., Lu, J.: Class semantics-based attention for action detection. In: ICCV (2021)
50. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: CVPR (2018)
51. Vishwakarma, S., Agrawal, A.: A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer **29**(10), 983–1009 (2013)
52. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: CVPR (2017)
53. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-tad: Sub-graph localization for temporal action detection. In: CVPR (2020)
54. Xu, Y., Zhang, C., Cheng, Z., Xie, J., Niu, Y., Pu, S., Wu, F.: Segregated temporal assembly recurrent networks for weakly supervised multiple action detection. In: AAAI (2019)
55. Yager, R.R., Liu, L.: Classic works of the Dempster-Shafer theory of belief functions, vol. 219. Springer (2008)
56. Yang, W., Zhang, T., Yu, X., Qi, T., Zhang, Y., Wu, F.: Uncertainty guided collaborative training for weakly supervised temporal action detection. In: CVPR (2021)
57. Yang, Z., Qin, J., Huang, D.: Acgnet: Action complement graph network for weakly-supervised temporal action localization. arXiv preprint arXiv:2112.10977 (2021)
58. Zhang, C., Cao, M., Yang, D., Chen, J., Zou, Y.: Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In: CVPR (2021)
59. Zhang, C., Xu, Y., Cheng, Z., Niu, Y., Pu, S., Wu, F., Zou, F.: Adversarial seeded sequence growing for weakly-supervised temporal action localization. In: ACM MM (2019)
60. Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y., Tian, Q.: Bottom-up temporal action localization with mutual regularization. In: ECCV. Springer (2020)
61. Zhong, J.X., Li, N., Kong, W., Zhang, T., Li, T.H., Li, G.: Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In: ACM MM (2018)
62. Zhu, Z., Tang, W., Wang, L., Zheng, N., Hua, G.: Enriching local and global contexts for temporal action localization. In: ICCV (2021)