Cascade Evidential Learning for Open-world Weakly-supervised Temporal Action Localization

Mengyuan Chen[1,2], Junyu Gao[1,2], and Changsheng Xu[1,2,3]
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Peng Cheng Laboratory, ShenZhen, China
chenmengyuan2021@ia.ac.cn; {junyu.gao, csxu}@nlpr.ia.ac.cn

Abstract

Targeting at recognizing and localizing action instances with only video-level labels during training, Weakly-supervised Temporal Action Localization (WTAL) has achieved significant progress in recent years. However, living in the dynamically changing open world where unknown actions constantly spring up, the closed-set assumption of existing WTAL methods is invalid. Compared with traditional open-set recognition tasks, Open-world WTAL (OWTAL) is challenging since not only are the annotations of unknown samples unavailable, but also the fine-grained annotations of known action instances can only be inferred ambiguously from the video category labels. To address this problem, we propose a Cascade Evidential Learning framework at an evidence level, which targets at OWTAL for the first time. Our method jointly leverages multi-scale temporal contexts and knowledge-guided prototype information to progressively collect cascade and enhanced evidence for known action, unknown action, and background separation. Extensive experiments conducted on THUMOS-14 and ActivityNet-v1.3 verify the effectiveness of our method. Besides the classification metrics adopted by previous open-set recognition methods, we also evaluate our method on localization metrics which are more reasonable for OWTAL.

1. Introduction

Targeting at recognizing and localizing action instances with only video-level labels during training, Weakly-supervised Temporal Action Localization (WTAL) has attracted increasing attention from both academia and industry [9, 11, 18, 19, 37, 43]. Unlike fully-supervised TAL, WTAL only requires video-level action labels during training. However, the closed-set assumption of existing WTAL methods is invalid in the dynamically changing real world, since with the development of society never-before-seen human action categories are constantly emerging. Therefore, to address this problem, we consider a different WTAL setting in this work, termed as Open-world WTAL (OWTAL).

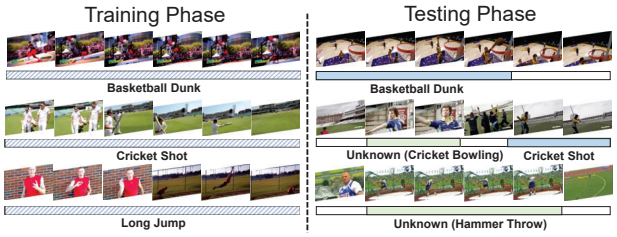

Figure 1. Illustration of the training and testing phases of OWTAL. With only video-level labels for training, OWTAL aims to localize both known and unknown action instances in testing videos.

Different from the traditional WTAL task, as shown in Figure 1, OWTAL allows testing videos to contain action instances of unknown categories, which have never appeared during training. Therefore, temporal boundries of both known and unknown action instances are expected to be predicted. Compared with its fully-supervised counterpart Open Set TAL [3], OWTAL is challenging in two aspects: **(1)** Ambiguity of annotations of closed-set (known) action instances. Previous works indicate that the closed-set and open-set performance are highly correlated [42]. However, under the OWTAL setting, not only are the annotations of unknown action instances unavailable, but also the fine-grained annotations of known ones can only be inferred ambiguously from the video category labels. During training, the known action instances that the model needs to focus on are prone to be disturbed by the background snippets, which hinders the learning of the closed-set actions, thus making it extremely difficult to differentiate the unknown actions, the known actions, and the background. **(2)** Lack of reasonable metrics. The traditional Open Set Recognition (OSR) aims

for classification while the goal of OWTAL is to perform localization instead, thus the classification metrics commonly adopted by OSR are not sufficient for OWTAL.

In order to alleviate the negative impact caused by the weak annotations of known action instances, we propose a Cascade Evidential Learning method for owtaL (CELL), which progressively collects cascaded evidence by considering both temporal contexts in multi-scale ranges and inter-video correlations under the guidance of prior knowledge. Since the goal of OWTAL is to locate the consecutive known/unknown action segments of various temporal lengths in open-world scenarios, perceiving temporal contexts in diverse ranges is essential. We argue that it is meaningful to endow individual snippet features with the ability of sensing multi-scale neighborhood video segments, and thus a Multi-scale Extended-range Perception module (Section 3.2) is designed to obtain more discriminative video features for initial evidence collection, taking advantage of the temporal contexts. Due to the large intra-action variation in visual patterns and the lack of prior knowledge guidance, the known action instances which visually deviate from the common ones are likely to be misidentified with the initial evidence collected from individual videos. Therefore, we design a Knowledge-guided Bipolar Prototype Learning strategy (Section 3.3), where a semantic relation graph is constructed to provide prior knowledge guidance for the bipolar prototype learning among videos, thus perceiving the open-world more comprehensively. We use this strategy to generate a series of evidence calibration factors for further cascade evidence enhancement. Finally, a Cascade Evidence Enhancement module (Section 3.4) is designed for enhancing the initial evidence with the calibration factors, and the uncertainty estimated from the cascaded evidence is used for the known/unknown judgment. Extensive experiments conducted on THUMOS-14 and ActivityNet verify the effectiveness. Besides the various classification metrics adopted by previous works, we also evaluate our method on localization metrics which are more in line with the needs of real-world applications.

To summarize, our contribution is threefold:

- To tackle the unique challenges of OWTAL, we propose a cascade evidential learning framework, which progressively collects comprehensive evidence for known action, unknown action, and background separation. Localization metrics which meet the needs of the real-world more closely are adopted for evaluation.
- To achieve OWTAL without fine-grained annotations, the proposed CELL jointly leverages multi-scale temporal contexts and knowledge-guided prototype information during the evidence cascade learning process.
- We conduct comprehensive experiments on two popular WTAL benchmarks, THUMOS-14 and ActivityNet-v1.3, and achieve significant performance improvement over various baselines. Experiments show that CELL enables existing methods to well adapt to the more practical open-world settings.

2. Related Work

Weakly-supervised Temporal Action Localization. Originated from UntrimmedNet [43], the pioneer work to utilize video-level action category annotations as the weak supervision for TAL, existing WTAL methods commonly adopt a multiple instance learning strategy and can be roughly categorized into erasing-based, attention-based and uncertainty-based methods. Erasing-based mthods [38, 49, 50] erase the most discriminative segments to mitigate the single snippet cheating issue [48]. Attention-based methods [10, 19, 21, 29, 33, 36] employ the attention mechanism to select snippets most likely to be the foreground with activation scores. The newly-arising uncertainty-based methods [9, 24, 46] utilize uncertainty to address the inevitable action-background ambuiguity caused by the weakly-supervised setting. However, all existing WTAL methods are rooted in the closed-set assumption, which impedes their application to real-world scenarios. In contrast, we address the WTAL task under the open-world setting by collecting reliable cascaded evidence, which has not been explored in prior work.

Open Set Recognition. Open Set Recognition (OSR) aims to recognize known classes and reject the unknown. [34] first formalize the OSR problem and provide a basic framework. [4] propose the first DNN-based OSR method Openmax, which rejects unknown classes by modeling the distance of activation vectors with Extreme Value Theory (EVT). GAN-based methods are also developed towards this task. [30] generate samples similar to the training data but not belonging to the known classes, and then utilize the generated unknown class samples to train an open-set classifier. Other approaches include prototype-based methods [7, 8], which reject unknown classes by calculating the maximum distance between the input sample and the learned closed-set prototypes, and reconstruction-based methods [31, 40, 47] utilize the reconstruction error in the test phase as an open-set indicator. Some works [28, 51] attempt to seek assistance from external knowledge, but few of them have explicitly modeled unknown class information in an end-to-end framework.

Most existing approaches focus on open-set recognition tasks while few works attend to temporal localization-related vision field. [3] formalize the open-set temporal action localization problem and propose an OpenTAL framework. However, its fully-supervised setting requires large amounts of fine-grained labels whose annotation is time-consuming, error-prone, and costly. Besides, the evaluation metrics in [3] are mainly classification ones. In contrast, in this paper, we address the challenging OWTAL problem which also abandons the closed-set assumption but only re-

quires easily available video-level class labels, and employ comprehensive localization-related metrics for evaluation.

Evidential Deep Learning. Based on Dempster-Shafer Theory (DST) [45] and Subjective Logic theory (SL) [22], Evidential Deep Learning (EDL) allows uncertainty estimation in a single forward pass [41] by collecting scalar evidence for each category and parameterizing a Dirichlet distribution, which models the distribution of classification probabilities, over the collected evidence. As a newly arising trustworthy method, EDL has achieved remarkable progress in various computer vision tasks, including action localization [9], image classification [35], regression [1], multi-view classification [17, 27], out-of-distribution detection [20], long-tail learning [25], and open-set action recognition and localization [2, 3]. However, current EDL algorithms only focus on collecting evidence from individual samples, while they neglect the complementary role that the potential correlations among samples can play in evidence collection. To address this issue, in this work we propose a cascade evidence calibration paradigm which effectively enhances the collected evidence.

3. Our Approach

3.1. Notations and Preliminaries

Problem formulation of OWTAL. OWTAL targets at detecting known/unknown action instances in untrimmed videos. Formally, we are given a training set $\{\mathcal{V}, \boldsymbol{y}\}$ with N training videos, where $\mathcal{V}$ denotes an untrimmed video, and $\boldsymbol{y} \in \mathbb{R}^C$ is its multi-hot label indicating the action categories that the action instances in this video belong to, where C is the number of known action categories. During testing, the goal is to use the learned model to output a set of quadruplets $\{c_r, t_r^s, t_r^e, \gamma_r\}_{r=1}^R$, where R is the number of action instances in $\mathcal{V}$, c_r, t_r^s, t_r^e and γ_r denote the action category, the start and end timestamps, and the confidence score, respectively. $c_r \in \{1, ..., C+1\}$, where $C+1$ denotes the unknown action category.

Feature extraction. Following previous works [10, 46], we split the untrimmed video $\mathcal{V}$ into T non-overlapping 16-frame snippets and then use pre-trained networks, *e.g.*, the I3D model [23], to extract RGB and optical flow features. Then, the above features are concatenated and then fed into a fusion module, *e.g.*, convolutional layers [19, 33], to obtain the snippet-wise feature $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_T] \in \mathbb{R}^{T \times D}$, where D is the feature dimension.

3.2. Multi-scale Extended-range Perception for Initial Evidence Collection

A video may contain several known/unknown action instances with various temporal lengths in the open-world. Since the goal of OWTAL is to locate these consecutive video segments, perceiving temporal contexts in diverse ranges is essential. We argue that it is meaningful to endow individual snippet features with the ability of sensing multi-scale neighborhood video segments. As shown in Figure 2, we design a Multi-scale Extended-range Perception module (MEP) to obtain more discriminative video features, taking advantage of the temporal contexts. Note that some anchor-based temporal modeling methods [6, 12, 14, 44] also leverage multi-scale temporal information. However, their goal is to obtain the fused temporal feature for proposal generation, while the MEP aims to enhance individual snippet features for initial evidence collection.

In MEP, each snippet t is assigned with a series of t-centered multi-scale video segments, whose context-aware features can contribute to the extension of snippet-specific perception ranges. Here, the temporal boundaries of the t-centered video segments can be denoted as set $\Omega_t = \{(s_t^m, e_t^m) | m \in \mathcal{M}\}$, where $\mathcal{M}$ is a pre-defined length set, $s_t^m = \max(0, t-m)$, and $e_t^m = \min(T, t+m)$. For brevity, we denote the temporal average of the feature sequence of the extended video segment (s_t^m, e_t^m) as $\boldsymbol{f}_t^m \in \mathbb{R}^D$, which can be given by:

$$\boldsymbol{f}_t^m = \frac{1}{e_t^m - s_t^m + 1} \sum_{s_t^m \le i \le e_t^m} \boldsymbol{x}_i. \tag{1}$$

To extend the perception range of snippet t, we perform feature enhancement by fusing $\boldsymbol{x}_t$ with $\boldsymbol{f}_t^m$, during which the weight of $\boldsymbol{f}_t^m$ is positively correlated with the similarity between them. Such operation is motivated by the simple expectation that a snippet-level feature would absorb meaningful contextual information mainly from extended video segments that are highly relevant to it. Therefore, the snippet-level feature with extended-range perception, denoted as $\tilde{\boldsymbol{F}} = [\tilde{\boldsymbol{f}}_1, \ldots, \tilde{\boldsymbol{f}}_T]$, can be formulated as:

$$\tilde{\boldsymbol{f}}_t = (1 - \alpha_t)\varphi_1(\boldsymbol{x}_t) + \alpha_t \sum_{m \in \mathcal{M}} \delta(\omega_t^m)\varphi_2(\boldsymbol{f}_t^m), \tag{2}$$

where $\delta(\cdot)$ denotes the Softmax operation with a temperature factor, $\varphi_{1,2}$ are fully-connected layers for feature embedding, $\alpha_t \in [0, 1]$ is the fusion weight representing the scaled average of the cosine similarities ω_t^m between $\boldsymbol{x}_t$ and $\boldsymbol{f}_t^m$, and α_t and ω_t^m can be calculated by:

$$\alpha_t = \frac{1}{2|\mathcal{M}|} \sum_{m \in \mathcal{M}} (\omega_t^m + 1), \omega_t^m = \cos\left(\frac{\boldsymbol{x}_t}{|\boldsymbol{x}_t|}, \frac{\boldsymbol{f}_t^m}{|\boldsymbol{f}_t^m|}\right). \tag{3}$$

Thereafter, similar to previous methods [10, 19], we utilize an attention module to predict an attention score sequence $\boldsymbol{A} = [A_1, ..., A_T] \in \mathbb{R}^T$ for the input video $\mathcal{V}$, representing the probabilities of snippets belonging to the foreground, then select the top-ranked snippets. Specifically, for the input video we obtain the set Θ of snippets with the top-k attention scores by the following equation:

$$\Theta = \arg\max_{\Theta} \sum_{t \in \Theta, |\Theta| = k} A_t, \quad \Theta \subset \{1, 2, ..., T\}, \tag{4}$$

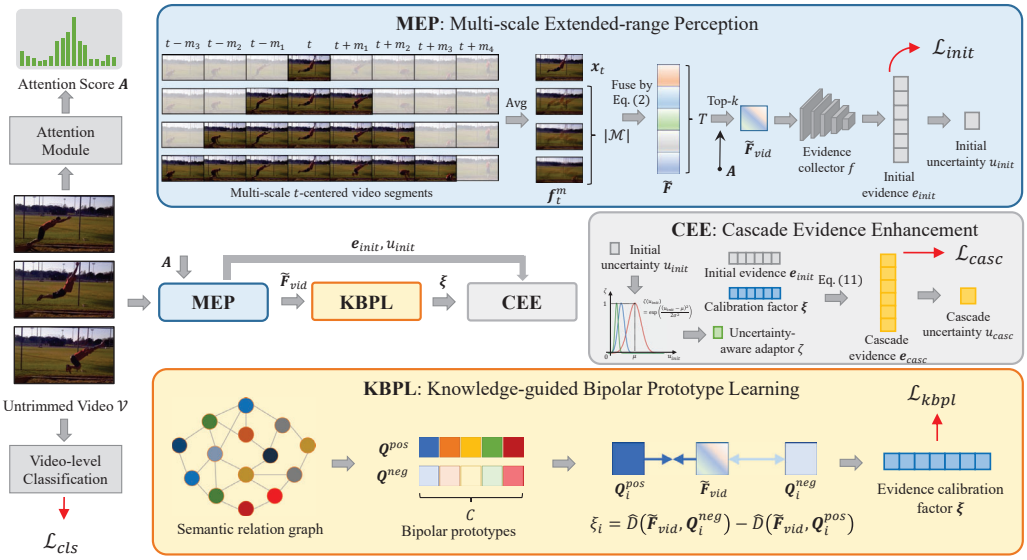

Figure 2. Overall framework of the proposed CELL for OWTAL. Firstly, a Multi-scale Extended-range Perception (MEP) module is designed to collect the initial evidence by perceiving temporal contexts in diverse ranges. Then we propose a Knowledge-guided Bipolar Prototype Learning (KBPL) module to obtain evidence calibration factors from inter-video correlations and prior knowledge. Finally, a Cascade Evidence Enhancement (CEE) module is designed for enhancing the initial evidence with the calibration factors, and the uncertainty derived from the cascaded evidence is used for the known/unknown judgment.

where $k = \lceil T/r \rceil$, r is a scaling factor. According to the snippet index in Θ, we select them from $\tilde{\boldsymbol{F}}$ and calculate the average as the video-level feature, denoted as $\tilde{\boldsymbol{F}}_{vid}$.

With the multi-scale-context-aware video-level feature, we feed $\tilde{\boldsymbol{F}}_{vid}$ into an evidence collector f, a DNN parameterized by $\boldsymbol{\theta}$, to collect initial evidence $\boldsymbol{e}_{init} \in \mathbb{R}^C$:

$$\boldsymbol{e}_{init} = g(f(\tilde{\boldsymbol{F}}_{vid}; \boldsymbol{\theta})), \tag{5}$$

where g denotes a scale function, *e.g.*, ReLU, Softplus or Exp, to ensure the collected evidence non-negative.

Based on Subjective Logic theory [22], Evidential Deep Learning (EDL) allows uncertainty estimation in a single forward pass [41] by collecting evidence of each category and parameterizing a Dirichlet distribution, which models the distribution of class probabilities, over the collected evidence. Our collected initial evidence $\boldsymbol{e}_{init}$ can be optimized by the following loss function:

$$\mathcal{L}_{init} = \sum_{i=1}^{C} y_i (\log S_{init} - \log \alpha_{init,i}), \tag{6}$$

where $S_{init} = \sum_i^C \alpha_{init,i}$, $\alpha_{init,i} = e_{init,i} + 1$, and Eq. (6) is actually the deformation of the negative logarithm of the marginal likelihood in the EDL paradigm [35]. Besides, the video-level classification uncertainty u_{init} can be derived by the EDL theory as $u_{init} = C/S_{init}$. Note that uncertainty u_{init} is inversely proportional to the total evidence of all closed-set categories, thus it reflects the probability that the video contains unknown actions.

3.3. Knowledge-guided Bipolar Prototype Learning for Evidence Calibration Factors

Due to the differences in action components, scenes, shooting angles and other aspects, videos tend to keep a large intra-action variation in visual patterns, and the known action instances which visually deviate from the majority are likely to be mistakenly identified as unknown ones. We argue that the initial evidence collected by a single video is insufficient for handling this issue. Therefore, we resort to the bipolar prototype learning strategy, in which both positive and negative prototypes are set for each category, to explore inter-video correlations for mining the intrinsic information of actions and eliminating the interference caused by background snippets. However, even if the intra-action visual variance is mitigated to some extent, it is still challenging to effectively detect the unknown action instances in the open-world environment for the lack of prior knowledge. To tackle the problem, a semantic relation graph is constructed to simulate the open world, so as to provide prior knowledge guidance for the bipolar prototype learning among videos, finally obtaining more accurate evidence calibration factors for further cascade evidence enhancement.

Based on the above observations, as shown in Figure 2, we propose a Knowledge-guided Bipolar Prototype Learning module (KBPL) to obtain evidence calibration factors from inter-video correlations and prior knowledge. Firstly, we construct a semantic relation graph with concepts related to known categories as vertice and their relations as edges. Specifically, the vertice set $\mathcal{N}$ includes concepts of known

classes, denoted as $\mathcal{N}_{kno}$, and the neighbor action concepts in ConceptNet [39] with the top-p relation strength, denoted as $\mathcal{N}_{nei}$, thus $\mathcal{N} = \mathcal{N}_{kno} \bigcup \mathcal{N}_{nei}$. Following a standard graph construction operation [13, 15], edges are connected in our graph if the relation strength between two concepts (vertice) is larger than a threshold. The features of vertice in $\mathcal{N}$ are initialized by the Glove-300 [32] embedding vectors, denoted as $\{\boldsymbol{n}_j\}_{j=1}^{|\mathcal{N}|}$, and then updated by:

$$\tilde{\boldsymbol{n}}_j = h_2(\underset{z\in\{1,2,3\}}{\text{Concat}}(\sum_{l\in\mathcal{S}_{j,z}} \delta(\eta_{j,l}) h_1(\boldsymbol{n}_l))) + h_3(\boldsymbol{n}_j), \quad (7)$$

where $h_{1,2,3}$ denote fully-connected layers, $\delta(\cdot)$ denotes the Softmax operation, $\mathcal{S}_{j,z}$ denotes the index set of the z-hop neighbor vertice of node j, and $\eta_{j,l}$ represents the cosine similarity of $\boldsymbol{n}_l$ and $\boldsymbol{n}_j$. Although many other complicated and advanced graph network methods can be adopted, we select the direct update manner here for simplicity.

With the concept relation graph, for category i, we use the updated representation of its corresponding vertex to generate the positive prototype, *i.e.* $\boldsymbol{Q}_i^{pos} = \phi_1(\boldsymbol{n}_i)$, and the average updated representaions of its neighbor vertice are adopted for producing the negative one, *i.e.* $\boldsymbol{Q}_i^{neg} = \phi_2(\sum_{j\in\mathcal{S}_i} \boldsymbol{n}_j/|\mathcal{S}_i|)$, where $\phi_{1,2}$ are fully-connected layers for prototype embedding. Then, for video $\mathcal{V}_b$ in the mini-batch, where $b = 1, ..., B$, and B is the mini-batch size, we define the calibration factor $\xi_{b,i}$ by simultaneously considering the distances between the video-level feature $\tilde{\boldsymbol{F}}_{vid,b}$ and the bipolar prototypes $\boldsymbol{Q}_i^{pos}$ and $\boldsymbol{Q}_i^{neg}$ as:

$$\xi_{b,i} = \hat{D}(\tilde{\boldsymbol{F}}_{vid,b}, \boldsymbol{Q}_i^{neg}) - \hat{D}(\tilde{\boldsymbol{F}}_{vid,b}, \boldsymbol{Q}_i^{pos}), \quad (8)$$

where $\hat{D}$ is a distance calculation method, which can be Euclidean distance or negative cosine similarity.

To guide the learning of prototypes by inter-video correlations, for each category, we pull close the videos belonging to the category to the corresponding positive prototype, and push away the videos of other classes in the same mini-batch, while opposite operations are performed for the negative prototype. Specifically, the loss function $\mathcal{L}_{kbpl}$ guiding the pos-neg prototype learning is designed as:

$$\mathcal{L}_{kbpl} = -\frac{1}{C}\sum_{i=1}^{C} \mathbb{I}(\mathcal{B}_i \neq \emptyset) \log\left(\frac{\sum_{b\in\mathcal{B}_i}\exp(\xi_{b,i})}{\sum_{b=1}^{B}\exp(\xi_{b,i})}\right), \quad (9)$$

where $\mathbb{I}(\cdot)$ is an indicator function, $\mathcal{B}_i = \{b|y_{b,i} = 1\}$, and $\boldsymbol{y}_b$ is the ground-truth label of the video $\mathcal{V}_b$.

3.4. Cascade Evidence Enhancement

As shown in Figure 2, we then perform cascade evidence enhancement, which scales the initial evidence collected from the MEP module by the learned calibration factors. As we stated above, the uncertainty u_{init} represents the probability that a video contains unknown actions. For example, given a test video, when u_{init} is close to the extreme values (close to zero for known actions and close to the maximum for unknown ones), we can recognize known or unknown classes with high confidence [35], otherwise the model has relatively low confidence in the known/unknown judgment. Specifically, we design an uncertainty-aware adaptor ζ as a Gaussian-like estimation function for evidence cascade:

$$\zeta(u_{init}; \mu, \sigma) = \exp\left(-\frac{(u_{init}-\mu)^2}{2\sigma^2}\right), \quad (10)$$

where μ is set to the value beyond the u_{init} of 95% training videos, and σ is a learnable scalar parameter which controls the shape of the Gaussian-like function.

If the above uncertainty-aware adaptor ζ of an input video is high, it means it is difficult for the model to perform known/unknown judgment by only considering the initial evidence collected inside the video, thus it is necessary to employ the information provided by the inter-video correlations to carry out evidence calibration. Specifically, the cascade evidence calibration process can be formulated as:

$$e_{casc,i} = (1 + \zeta(u_{init})\tanh(\xi_i))e_{init,i}, \quad (11)$$

where the $\tanh(\cdot)$ function ensures the scaling range of $e_{init,i}$ to be $(0, 2)$ for a stable learning, and $\boldsymbol{e}_{casc} = [e_{casc,1}, ..., e_{casc,C}] \in \mathbb{R}^C$ is the enhanced cascaded evidence. The subscript b (video index) is omitted in this equation for simplicity. $\boldsymbol{e}_{casc}$ is then optimized by an EDL loss $\mathcal{L}_{casc}$, whose form is identical with Eq. (6), and the cascaded video-level uncertainty u_{casc} for known/unknown judgment can be obtained in the same way with u_{init}.

3.5. Training and Inference

Training. By combining all the optimization objectives introduced above, we obtain the final loss function as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{init} + \mathcal{L}_{kbpl} + \mathcal{L}_{casc}, \quad (12)$$

where $\mathcal{L}_{cls}$ is the standard video-level classification loss, *e.g.*, MIL loss [10, 19], for optimizing the backbone.

Inference. For each video, we predict its action categories set Ψ (including the *Unknown* class if exist) by Algorithm 1, where the closed-set video-level classification score $\boldsymbol{P} = [P_1, ..., P_C]$ is obtained by averaging the classification activation sequence (CAS) of snippets in Θ. Then we use a threshold strategy to obtain action snippet candidates following the standard process [18, 19]. Finally, for each action class in Ψ we group continuous snippets into action proposals, whose confidence scores γ are estimated according to the CAS scores of the corresponding class (for *Known* classes) or attention scores (for the *Unknown* class), and perform non-maximum-suppression (NMS) to remove duplicated proposals.

Table 1. OWTAL localization results on THUMOS-14 evaluated by both top-K mAP@Avg and traditional mAP@Avg. The averages of the mAP of *Known* classes and the AP of the *Unknown* class are reported. The values are averaged on t-IoU [0.1:0.1:0.5] and [0.1:0.1:0.7].

Methods	Top-K mAP@Avg(%)										mAP@Avg(%)	
	Top-5		Top-10		Top-20		Top-50		Top-100			
	0.1-0.5	0.1-0.7	0.1-0.5	0.1-0.7	0.1-0.5	0.1-0.7	0.1-0.5	0.1-0.7	0.1-0.5	0.1-0.7	0.1-0.5	0.1-0.7
ASM-Loc + Trivial	4.3	3.5	7.8	6.3	13.3	10.7	21.1	16.9	26.1	20.9	29.5	23.6
ASM-Loc + SoftMax	10.9	8.9	16.9	13.5	23.4	18.7	28.8	23.0	29.8	23.8	30.1	24.0
ASM-Loc + OpenMax	10.1	8.1	15.3	12.1	21.2	16.7	26.1	20.6	27.3	21.5	27.6	21.7
ASM-Loc + ARPL	10.0	8.2	16.4	13.4	23.6	19.3	30.4	24.8	32.0	26.0	32.4	26.4
ASM-Loc + EDL	**11.3**	9.2	17.4	14.0	24.2	19.4	30.5	24.3	31.9	25.4	32.2	25.7
ASM-Loc + CELL(Ours)	11.2	**9.3**	**18.0**	**14.6**	**25.5**	**20.7**	**32.4**	**26.4**	**34.1**	**27.8**	**34.7**	**28.1**
CO2-Net + Trivial	5.5	4.4	9.5	7.7	16.7	13.5	25.9	20.9	30.9	25.0	34.4	27.9
CO2-Net + SoftMax	11.3	9.1	17.8	14.3	25.1	20.2	32.2	26.0	33.7	27.3	34.2	27.8
CO2-Net + OpenMax	10.3	8.4	16.3	13.2	23.0	18.6	29.1	23.5	30.4	24.7	30.8	25.0
CO2-Net + ARPL	11.6	9.5	18.3	14.9	25.7	20.9	33.3	27.1	35.1	28.7	35.7	29.2
CO2-Net + EDL	11.2	9.1	17.6	14.3	24.8	20.0	32.2	26.0	34.0	27.5	34.6	28.1
CO2-Net + CELL(Ours)	**12.6**	**10.3**	**20.1**	**16.4**	**28.1**	**23.0**	**36.9**	**30.3**	**38.9**	**31.8**	**39.5**	**32.3**

Table 2. OWTAL localization results on THUMOS-14 evaluated by both top-K mAP@Avg and traditional mAP@Avg. The mAP of *Known* classes (K) and the AP of the *Unknown* class (U) averaged on t-IoU thresholds [0.1:0.1:0.7] are reported respectively. The video-level known-unknown classification accuracy (v-Acc(%)) is also presented for reference.

Methods	Top-K mAP@Avg(%)										mAP@Avg(%)		v-Acc(%)
	Top-5		Top-10		Top-20		Top-50		Top-100				
	K	U	K	U	K	U	K	U	K	U	K	U	
CO2-Net + Trivial	4.4	4.4	7.8	7.6	15.1	11.9	23.8	18.0	30.5	19.4	36.2	19.6	28.1
CO2-Net + SoftMax	12.6	5.6	19.3	9.3	26.4	13.9	32.7	19.2	34.5	20.0	35.6	20.0	71.2
CO2-Net + OpenMax	11.9	4.9	18.4	8.0	25.1	12.0	31.0	16.1	32.7	16.7	33.6	16.5	70.5
CO2-Net + ARPL	12.3	6.7	19.4	10.4	26.6	15.2	32.6	21.7	34.6	22.7	35.8	22.6	73.0
CO2-Net + EDL	12.4	5.7	19.1	9.4	26.3	13.8	32.7	19.4	34.8	20.2	35.9	20.3	70.9
CO2-Net + CELL(Ours)	**13.3**	**7.3**	**20.5**	**12.2**	**27.6**	**18.3**	**34.0**	**26.6**	**35.9**	**27.6**	**36.8**	**27.8**	**74.9**

4. Experimental Results

4.1. Experimental Setup

Datasets. We conduct experiments on two popular benchmarks for OWTAL, THUMOS-14 and ActivityNet-v1.3. THUMOS-14 contains 200 validation videos and 213 test videos annotated from 20 action categories, and each video contains 15.4 action instances on average. ActivityNet-v1.3 contains 10,024 training videos and 4,926 validation videos from 200 action categories, and each video contains 1.6 action instances on average. Following the pioneer work [3], we randomly remove 1/4 action categories of THUMOS-14 training set and keep the entire THUMOS-14 testing set to enable open-set evaluation, and such random removal is repeated three times. To further increase the openness and verify the applicability of our proposed method for open-world scenarios, we train our model on the known split of THUMOS-14 and test on the testing set of ActivityNet-v1.3. Note that since ActivityNet-v1.3 covers most categories in THUMOS-14, we manually remove 14 action categories which are semantically overlapping with the THUMOS-14 as previous work did.

Evaluation Metrics. We adopt three sets of metrics to measure the performance of the OWTAL task. (1) **Vanilla mAP**. To evaluate the localization performance, following previous approaches [10,19,33], we employ the mean Average Precision (mAP) under different temporal Intersection over Union (t-IoU) thresholds as metrics. The t-IoU thresholds for THUMOS-14 is [0.1:0.1:0.7] and for ActivityNet-v1.3 is [0.5:0.05:0.95]. (2) **Top-K mAP**. In our experiments, we observe that it only brings slight performance loss on the traditional mAP for models to generate a large number of redundant proposals. However, in real-world applications our goal is to provide users with accurate action proposals of a limited number, rather than an extreme large-scale proposal set consisting of numerous redundant proposals, although the set may contain even more accurate predictions. Inspired by the above observation, we design a

Algorithm 1 Action Categories Inference Procedure

Input: Untrimmed testing video $\mathcal{V}$.
Require: Trained CELL model.
Require: Threshold τ obtained from training data (Please refer to **Implementation Details**).
Output: Set Ψ of action categories in $\mathcal{V}$.

Predict the closed-set classification score $\boldsymbol{P}$ and cascaded video-level uncertainty u_{casc} by CELL.
if $u_{casc} < \tau$ **then**
  $\Psi = \{i | P_i > 0.2\}$ ▷ Only Known Classes
  if $\Psi = \varnothing$ **then**
    $\Psi = \arg\max_i P_i$ ▷ Only Known Classes
  end if
else if $\arg\max_i P_i > 0.5$ **then**
  $\Psi = \{\arg\max_i P_i, C+1\}$
    ▷ Both Known and Unknown Classes
else
  $\Psi = \{C+1\}$ ▷ Only Unknown Classes
end if
return Ψ

variant metric of the vanilla mAP to further meet the needs of real-world applications, termed as top-K mAP, which restricts the maximum number of generated proposals by selecting proposals with top-K confidence scores. When K is not restricted, the top-K mAP degrades to the traditional mAP metric. In our experiments K is set to 5, 10, 20, 50, 100 for THUMOS-14 and 1, 3, 5 for ActivityNet-v1.3, respectively. (3) **Classification metrics**. For known/unknown classification, following [3], four classification metrics including the False Alarm Rate at True Positive Rate of 95% (FAR@95) (smaller value indicates better performance), the Area Under the Receiver Operating Characteristic (AUROC) curve, the Area Under the Precision-Recall (AUPR), and the Open Set Detection Rate (OSDR) are also employed. To validate the effectiveness of Algorithm 1, we additionally adopt the video-level known-unknown classification accuracy (v-Acc(%)) metric for evaluation. Note that a testing video may belong to three types which have: (i) only known actions, (ii) only unknown actions, or (iii) both known and unknown actions.

Implementation Details. We adopt CO2-Net [19] and ASM-Loc [18] as our backbone networks. Following existing methods, we use I3D [5] model pretrained on the Kinetics [23] dataset to extract both the RGB and optical flow features. The feature dimension D is 2048, the evidence scale function g in Eq. (6) is Exp, and the distance calculation method $\hat{D}$ in Eq. (8) is negative cosine similarity. The number of the sampled snippets T for THUMOS-14 and ActivityNet-v1.3 is 500 and 160, and the scaling factor r is 7. ξ in Eq. (10) is updated for every 25 epochs. The batch size is set to 10 and 16, and the learning rate is 5e-5 and 1e-4 in CO2-Net and ASM-Loc, respectively. Following previous works [2, 40], we determine the threshold τ of uncertainty u_{casc} in Algorithm 1 by ensuring 95% training videos to be recognized as known. Our model is implemented with Python 3.7 and PyTorch 1.11.0. All experiments are conducted on a single RTX3090 GPU.

4.2. Comparision with State-of-the-art Methods

Our proposed CELL method is compared with the following baselines based on the ASM-Loc [18] and the CO2-Net [19]: (1) **Trivial**: assuming that all testing videos contain both known and unknown actions without any known/unknown judgment. (2) **SoftMax**: utilizing the maximum softmax probability to identify the unknown. (3) **OpenMax**: appending the unknown score to the softmax scores by OpenMax [4] in testing. (4) **ARPL**: using the Adversarial Reciprocal Points Learning (ARPL) [7] to identify the unknown. (5) **EDL**: vanilla EDL is used to replace the softmax classification head for uncertainty estimation.

Comparison results on THUMOS-14. Table 1 and Table 2 represent the OWTAL localization results on the THUMOS-14 dataset. The results show that our proposed method consistently outperforms the baselines by large margins on all metrics. We notice that totally without any known/unknown judgment, the Trivial method shows comparable performance with the SoftMax baseline on the traditional mAP metric, for the embarrassing property that it only brings slight performance loss on the traditional mAP to generate numerous error redundant proposals, while on our proposed top-k mAP metrics the performance of the Trivial version shows a reasonable gap. The results also show that Openmax does not work well on the OWTAL task, which may lie in that Openmax often fails to recognize visually indistinguishable samples [16], especially in weakly-supervised video analysis field. Note that the other two SOTA approaches ARPL and EDL perform well but still far behind the proposed CELL.

Table 3 displays the known/unknown classification results evaluated on the metrics adopted by [3], showing that our method outperforms the EDL baseline and even achieves comparable results with fully-supervised methods on FAR@95, AUROC, and AUPR. We infer that the large gap between weakly-supervised methods and fully-supervised ones on OSDR is due to the difference between backbone networks. OSDR is defined as the area under the curve of Correct Detection Rate (CDR) and False Positive Rate (FPR), and the CDR indicates the fraction of known actions which are positively localized and correctly classified, while the FPR denotes the fraction of unknown actions that are positively localized but falsely classified. CO2-Net adopts the localization-by-classification strategy commonly used in WTAL method, thus the positively localized actions are usually accompanied by correct class predictions, while AFSD [26] adopts an anchor-free method whose localiza-

tion performance does not entirely rely on the classification results, thus resulting in the large gap on OSDR. Note that compared with our employed mAP-related metrics, OSDR is still a classification metric thus cannot fully exhibit the superiority of different models.

Table 3. OWTAL classification results on THUMOS-14 evaluated by FAR@95, AUROC, AUPR and OSDR. In this experiment we use CO2-Net as our backbone.

Supervision	Methods	FAR@95(↓)	AUROC	AUPR	OSDR
Fully	SoftMax	85.58	54.70	31.85	23.40
	OpenMax	90.34	53.26	33.17	13.66
	EDL	81.42	64.05	40.05	36.26
	OpenTAL	70.96	78.33	58.62	42.91
Weakly	CO2+EDL	87.37	67.11	44.47	63.56
	CELL(Ours)	**68.86**	**74.62**	**54.25**	**69.81**

Table 4. OWTAL localization results on ActivityNet-v1.3. The values are averaged on t-IoU thresholds [0.5:0.05:0.95]. The metric v-Acc(%) is also presented for reference.

Methods	Top-K mAP@Avg(%)			mAP@Avg(%)	v-Acc(%)
	Top1	Top3	Top5		
ASM-Loc + SoftMax	12.7	13.2	13.5	13.5	86.37
ASM-Loc + OpenMax	12.0	12.6	12.8	12.9	81.74
ASM-Loc + ARPL	17.0	17.6	17.8	18.0	96.81
ASM-Loc + EDL	16.0	16.7	16.9	17.0	93.38
ASM-Loc + CELL(Ours)	**17.9**	**18.4**	**18.6**	**18.9**	**97.83**

Table 5. Ablation study of our proposed CELL. MEP denotes the multi-scale extended-range perception module, KBPL represents the knowledge-guided bipolar prototype learning module, and UA denotes the uncertainty-aware adaptor ζ described in Eq. (10).

Exp	MEP	KBPL	UA	Top-k mAP@Avg(%)			mAP@Avg(%)
				Top-10	Top-20	Top-50	
1	✗	✗	✗	14.3	20.2	26.0	27.8
2	✓	✗	✗	15.3	21.2	26.9	29.0
3	✗	✓	✗	15.2	21.1	26.9	29.2
4	✓	✓	✗	16.0	22.4	29.4	31.6
5	✓	✓	✓	**16.4**	**23.0**	**30.3**	**32.3**

Comparison results on ActivityNet-v1.3. To further increase the openness and verify the applicability of our proposed method for open-world scenarios, we further train our model on the known splits of THUMOS-14 and test on ActivityNet-v1.3. Although the testing set only consists of unknown action categories, in order to ensure a fair comparison, the model is still tested according to Algorithm 1, without using any additional prior knowledge. As shown in Table 4, our proposed CELL obtains consistently favorable performance. Compared with EDL, CELL shows superior performance with an absolute gain of 1.9% in mAP@Avg.

4.3. Ablation Study

This section shows the effectiveness of our modules. In Table 5, the impact of progressively adding each component is presented, proving their contributions clearly. According to the cascade steps of our method, we perform ablation study on three components, Multi-scale Extended-range Perception module (MEP), Knowledge-guided Bipolar Prototype Learning module (KBPL), and the Uncertainty-aware Adaptor (UA) in CEE module. Note that without MEP we replace the fused feature by the average of snippet features in Θ, and without UA we simply set ζ in Eq. (11) to 1. As shown in Table 5, every step of our method brings effective performance improvement. Moreover, it is noteworthy that MEP perceiving temporal contexts in multi-scale ranges and KBPL leveraging inter-video correlations and prior knowledge can enhance and complement each other, thus significantly improving the performance.

5. Conclusions

Targeting at open-world weakly-supervised temporal action localization (OWTAL), we propose a cascade evidential learning framework, which entails three main components: (1) Multi-scale extended-range perception module perceiving temporal contexts in diverse ranges for collecting initial evidence; (2) Knowledge-guided bipolar prototype learning strategy exploring inter-video relations under the guidance of prior knowledge for seeking supplementary evidence support; (3) Cascade evidence enhancement for final evidence calibration. In our extensive experiments, CELL achieves state-of-the-art performance on various metrics. Several limitations of this work are noteworthy. Firstly, never-before-seen human action categories are constantly emerging, thus lifelong learning paradigm which can keep embracing new action categories is even more suitable for OWTAL. Secondly, the class annotation granularity of the training set will interfere with the unknown action identification in open-world scenarios. Specifically, if some fine-grained actions are not labeled during training, *e.g.*, the run-up movement in *HighJump* videos in THUMOS14, the model may misjudge such unknown actions as the background in testing. We assume the problem can be alleviated by training on large-scale datasets with fine-grained annotations. These directions are left as future work.

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, 62236008, 62102415, 61721004, 62072286, 62072455, 62002355, in part by Beijing Natural Science Foundation (L201001), and in part by Open Research Projects of Zhejiang Lab (NO.2022RC0AB02).

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *NeurIPS*, 2020. 3

[2] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *ICCV*, 2021. 3, 7

[3] Wentao Bao, Qi Yu, and Yu Kong. Opental: Towards open set temporal action localization. In *CVPR*, 2022. 1, 2, 3, 6, 7

[4] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016. 2, 7

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 7

[6] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, 2018. 3

[7] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *arXiv preprint arXiv:2103.00953*, 2021. 2, 7

[8] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*. Springer, 2020. 2

[9] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *ECCV*. Springer, 2022. 1, 2, 3

[10] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization. In *CVPR*, 2022. 2, 3, 5, 6

[11] Junyu Gao and Changsheng Xu. Fast video moment retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1523–1532, 2021. 1

[12] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4649–4659, 2019. 3

[13] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In *AAAI*, 2019. 5

[14] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Smart: Joint sampling and regression for visual tracking. *IEEE Transactions on Image Processing*, 28(8):3923–3935, 2019. 3

[15] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Learning to model relationships for zero-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3476–3491, 2021. 5

[16] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020. 7

[17] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *ICLR*, 2020. 3

[18] Bo He, Xitong Yang, Le Kang, Zhiyu Cheng, Xin Zhou, and Abhinav Shrivastava. Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In *CVPR*, 2022. 1, 5, 7

[19] Fa-Ting Hong, Jia-Chang Feng, Dan Xu, Ying Shan, and Wei-Shi Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In *ACM MM*, 2021. 1, 2, 3, 5, 6, 7

[20] Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen. Multidimensional uncertainty-aware evidential neural networks. In *AAAI*, 2021. 3

[21] Linjiang Huang, Liang Wang, and Hongsheng Li. Weakly supervised temporal action localization via representative snippet knowledge propagation. In *CVPR*, 2022. 2

[22] Audun Jsang. Subjective logic: A formalism for reasoning under uncertainty. *Springer Verlag*, 2016. 3, 4

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv:1705.06950*, 2017. 3, 7

[24] Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*, 2021. 2

[25] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. *arXiv preprint arXiv:2111.09030*, 2021. 3

[26] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In *CVPR*, 2021. 7

[27] Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. Trusted multi-view deep learning with opinion aggregation. In *AAAI*, 2022. 3

[28] Vincent Lonij, Ambrish Rawat, and Maria-Irina Nicolae. Open-world visual recognition using knowledge graphs. *arXiv preprint arXiv:1708.08310*, 2017. 2

[29] Wang Luo, Tianzhu Zhang, Wenfei Yang, Jingen Liu, Tao Mei, Feng Wu, and Yongdong Zhang. Action unit memory network for weakly supervised temporal action localization. In *CVPR*, 2021. 2

[30] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018. 2

[31] Poojan Oza and Vishal M Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *CVPR*, 2019. 2

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 5

[33] Sanqing Qu, Guang Chen, Zhijun Li, Lijun Zhang, Fan Lu, and Alois Knoll. Acm-net: Action context modeling network for weakly-supervised temporal action localization. *arXiv preprint arXiv:2104.02967*, 2021. 2, 3, 6

[34] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 2

[35] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, 2018. 3, 4, 5

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

[37] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In *ECCV*, 2018. 1

[38] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *ICCV*, 2017. 2

[39] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017. 5

[40] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. Conditional gaussian distribution learning for open set recognition. In *CVPR*, 2020. 2, 7

[41] Dennis Ulmer. A survey on evidential deep learning for single-pass uncertainty estimation. *arXiv preprint arXiv:2110.03051*, 2021. 3, 4

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2022. 1

[43] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017. 1, 2

[44] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 3

[45] Ronald R Yager and Liping Liu. *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer, 2008. 3

[46] Wenfei Yang, Tianzhu Zhang, Xiaoyuan Yu, Tian Qi, Yongdong Zhang, and Feng Wu. Uncertainty guided collaborative training for weakly supervised temporal action detection. In *CVPR*, 2021. 2, 3

[47] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *CVPR*, 2019. 2

[48] Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*, 2021. 2

[49] Chengwei Zhang, Yunlu Xu, Zhanzhan Cheng, Yi Niu, Shiliang Pu, Fei Wu, and Futai Zou. Adversarial seeded sequence growing for weakly-supervised temporal action localization. In *ACM MM*, 2019. 2

[50] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one collection: a weakly supervised temporal action detector. In *ACM MM*, 2018. 2

[51] Zhun Zhong, Linchao Zhu, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Openmix: Reviving known knowledge for discovering novel visual categories in an open world. In *CVPR*, 2021. 2