OpenTAL: Towards Open Set Temporal Action Localization

Wentao Bao, Qi Yu, Yu Kong

Rochester Institute of Technology, Rochester, NY 14623, USA

{wb6219, qi.yu, yu.kong}@rit.edu

Abstract

*Temporal Action Localization (TAL) has experienced remarkable success under the supervised learning paradigm. However, existing TAL methods are rooted in the closed set assumption, which cannot handle the inevitable unknown actions in open-world scenarios. In this paper, we, for the first time, step toward the Open Set TAL (OSTAL) problem and propose a general framework **OpenTAL** based on Evidential Deep Learning (EDL). Specifically, the OpenTAL consists of uncertainty-aware action classification, actionness prediction, and temporal location regression. With the proposed importance-balanced EDL method, classification uncertainty is learned by collecting categorical evidence majorly from important samples. To distinguish the unknown actions from background video frames, the actionness is learned by the positive-unlabeled learning. The classification uncertainty is further calibrated by leveraging the guidance from the temporal localization quality. The OpenTAL is general to enable existing TAL models for open set scenarios, and experimental results on THUMOS14 and ActivityNet1.3 benchmarks show the effectiveness of our method. The code and pre-trained models are released at* https://www.rit.edu/actionlab/opental.

1. Introduction

Temporal Action Localization (TAL) aims to temporally localize and recognize human actions in an untrimmed video. With the success of deep learning in video understanding [4,10,15,20,34] and object detection [3,9,49], TAL has experienced remarkable advance in recent years [12,36, 62,69]. However, these works are rooted in the closed set assumption that testing videos are assumed to contain only the pre-defined action categories, which is impractical in an open world where unknown human actions are inevitable to appear. In this paper, we for the first time step forward the Open Set Temporal Action Localization (OSTAL) problem.

OSTAL aims to not only temporally localize and recognize the known actions but also reject the localized unknown actions. As shown in Fig. 1, given an untrimmed
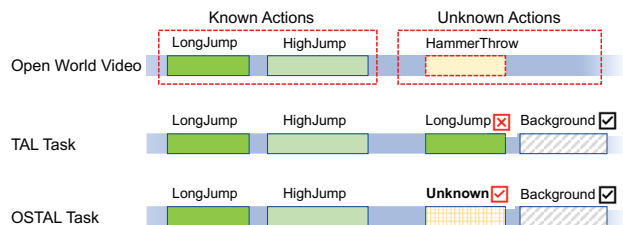

Figure 1. **OSTAL and TAL Tasks**. The OSTAL task is different from the TAL in that, there exist unknown actions in untrimmed open-world videos and the OSTAL models need to reject the positively localized action (e.g., *HammerThrow*) as the *Unknown*, rather than falsely assign it a known label such as the *LongJump*.

video (the top row) from open world, traditional TAL (the middle row) could falsely accept the unknown action clip *HammerThrow* as one of the known actions such as the *LongJump*, while the proposed OSTAL (the bottom row) could correctly reject the clip as the *Unknown*. Besides, both tasks need to differentiate between foreground actions and the *Backgrounds* which are purely background frames.

The proposed OSTAL task is fundamentally more challenging than both the TAL and the closely relevant open set recognition (OSR) [51] problems. On one hand, the recognition and localization of known actions become harder due to the mixture of background frames and the unknown foreground actions. Existing TAL methods typically assign the mixture with an non-informative *Background* label or a wrong action label, which are unable to differentiate between them. On the other hand, different to the OSR problem, rejecting an unknown action is conditioned on positively localizing a foreground action so that the localization quality is critical to the OSTAL.

To tackle these challenges, we propose a general framework OpenTAL by decoupling the overall OSTAL objective into three interconnected components: uncertainty-aware action classification, actionness prediction, and temporal location regression. In essence, the foreground actions are distinguished from the background by the actionness prediction and localized by the temporal localization, while the known and unknown foreground actions are discriminated

by the learned evidential uncertainty from the classification module. To achieve these goals, we propose three novel technical approaches as follows.

First, action classification is developed to recognize known actions and quantify the classification uncertainty by recent evidential deep learning (EDL) [1, 4, 53, 72]. To enable this module to learn from important samples, we propose an importance-balanced EDL method by leveraging the magnitude of EDL gradient and evidential features. Second, actionness prediction is to differentiate between foreground actions (positives) and background frames (negatives). In the open set setting, due to the mixture of unknown foreground actions (unlabeled) and background frames, learning from the labeled known actions and the mixture intrinsically reduces to a positive-unlabeled (PU) learning problem [5]. To this end, we propose a PU learning method by selecting the top negative samples from the mixture as the true negatives. Third, the temporal localization module is trained to not only localize the known actions but also calibrate the classification uncertainty. We propose an IoU-aware uncertainty calibration (IoUC) method by using the temporal Intersection-over-Union (IoU) as the localization quality to calibrate the uncertainty.

Based on the existing TAL datasets THUMOS14 [27] and ActivityNet1.3 [8], we set up a new benchmark to evaluate baselines and the proposed OpenTAL method for the OSTAL task, where the *Open Set Detection Rate* is introduced to comprehensively evaluate the OSTAL performance. Experimental results show significant superiority of our method and indicate large room for improvement in this direction. Our main contribution is threefold:

- To the best of our knowledge, this work is the first attempt on open set temporal action localization (OSTAL), which is fundamentally more challenging but highly valuable in open-world settings.

- We propose a general OpenTAL framework to address the unique challenges of OSTAL as compared with existing TAL and OSR problems. It is flexible to enable existing TAL models for open set scenarios.

- The proposed importance-balanced EDL, PU learning, and IoUC methods are found effective for OSTAL tasks based on the OpenTAL framework.

2. Related Work

Temporal Action Localization The goal of Temporal Action Localization (TAL) is to recognize and temporally localize all the action instances in an untrimmed video. Existing TAL methods fall into two dominant paradigms: one-stage and two-stage approaches. The two-stage approaches [40,57,61,62,69] generate class-agnostic temporal proposals [2, 24, 37, 38] at first and then perform the classification and boundary refinement of each proposal. The heuristic anchor design and the closed-set definition of the pre-trained proposal generation limit their applicability to the open-set problem. One stage methods [7, 36, 40, 65] do not rely on the action proposal generation and can be typically trained in an end-to-end manner. These methods obtain the temporal boundaries first based on frame-level features and then perform global reasoning by multi-stage refinement or modeling the temporal transitions. Recently, AFSD [36] is proposed following the anchor-free design without actionness and proposals, which is a lightweight and flexible framework. While a lot of recent methods focus on improving the proposal generation [2, 24, 37, 38, 71] or boundary refinement [36], a few focus on boosting the classification accuracy [55, 73, 75].

The above approaches assume that all of the action instances in untrimmed videos belong to pre-defined categories, which impedes their application to open-world scenarios. Though the open set is considered in [76], their method is designed for efficient annotation in few-shot learning tasks. In this paper, an OSTAL problem is formulated to handle the unknown actions in TAL applications.

Open Set Recognition Open set recognition (OSR) aims to recognize known classes and reject the unknown. The pioneering work by Scheirer *et al.* [51] formalized the definition of OSR and introduced an "one-vs-set" machine based on binary SVM, which inspired a line of SVM-based OSR methods [26, 30, 52]. Benefited by the deep neural networks (DNNs), Bendale *et al.* [6] proposed the first DNN-based OSR method OpenMax, which leverages Extreme Value Theory (EVT) to expand the K-class softmax classifier. Recently, Fang *et al.* [19] theoretically proved the learnability of OSR classifier and the generalization bound. Existing generative OSR methods [13, 18, 21, 33, 47, 68, 74] utilize GAN [23], generative causal model, or mixup augmentation to generate the samples of the unknown. From the reconstruction perspective, some literature [45, 58, 66] leverage VAE [31] or self-supervised learning to reconstruct the representation of known class data to identify the unknown. Prototype learning and metric learning methods [11, 13, 14, 56, 63, 64, 70] aim to identify the unknown by producing large distance to the prototype of known class data. Recently, uncertainty estimation methods [4, 42, 60] by probabilistic and evidential deep learning show promising results on OSR problems.

In this paper, we step further toward the OSTAL problem. We are aware of analogous extensions from OSR to open set object detection [16, 29, 41] and segmentation [25, 44, 48, 59]. However, it is the uniqueness of the localization in open-world that makes the OSTAL problem even more challenging and valuable in practice.
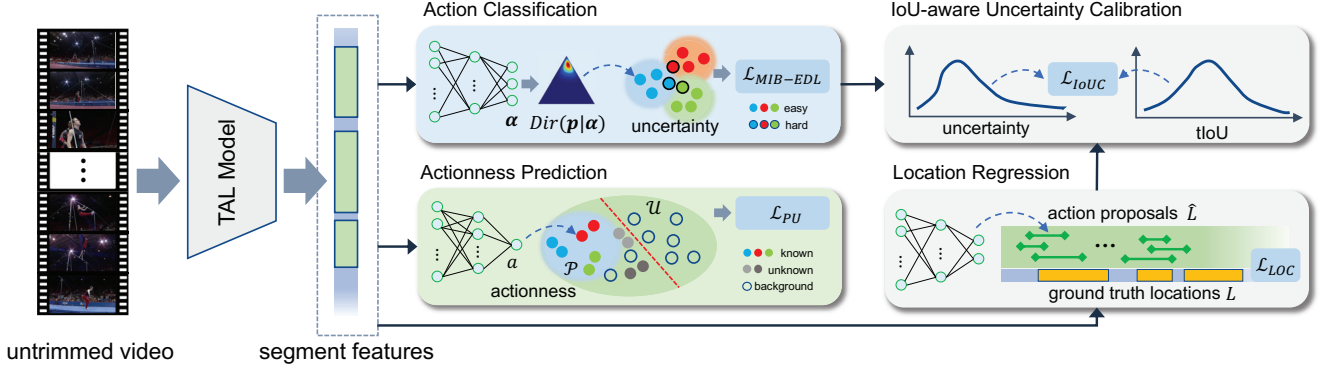
Figure 2. **The Proposed OpenTAL.** Given untrimmed videos as input, the OpenTAL method is developed on existing TAL models (such as AFSD [36]) toward the OSTAL scenario. It consists of action classification, actionness prediction, and location regression, which are learned by the proposed MIB-EDL loss (Eq. (5)), PU learning (Eq. (6)), and localization loss (Eq. (7)), respectively. Furthermore, the IoU-aware uncertainty calibration is proposed to calibrate the uncertainty estimation by considering localization quality (Eq. (8)). In inference, with a two-step decision procedure by leveraging the uncertainty and actionness, video actions from the known and unknown classes, as well as background frames can be distinguished in the OSTAL setting (see Algorithm 1).

3. Proposed Method

Setup Given an untrimmed video, the OSTAL task requires a model to localize all actions with temporal locations $l_i = (s_i, e_i)$, assign the actions with labels $y_i \in \{0, 1, \dots, K\}$ where $y_i = 0$ indicates the action consisting of background frames, and reject the actions from novel classes as the unknown. In the training, the model only has access to the video data and the annotations of known actions, while the annotations of unknown actions are not given. This setting is different from the OSR problem where both annotations and data of unknown classes are not given, because it is impractical in the TAL task to discard video segments of unknown actions.

Overview Fig. 2 shows an overview of the proposed OpenTAL. Given an untrimmed video, the features of action proposals are obtained from an existing TAL model such as the AFSD [36]. To fulfill OSTAL, we decouple the objective into three sub-tasks by a trident head, including action classification, actionness prediction, and location regression. The three branches are learned by multi-task loss functions, which will be introduced in detail.

Motivations Existing TAL models typically adopt a $(K+1)$-way action classification by assigning the background video frames with the $(K+1)$-th class *Background*. However, this paradigm is unable to handle the OSTAL case when unknown actions exist in the *Background* class.

To solve this problem, on one hand, one would attempt to append the K known classes with an additional *Unknown* category in an existing TAL system. However, this solution is practically infeasible under the OSTAL setting, because finding the video segments to train a classifier with the class

Unknown relies on the temporal boundary annotations of unknown actions, which are not available under our OSTAL setting. Though one could relax the OSTAL setting by providing temporal annotations of the unknowns in training, learning a $(K+1)$-way classifier is nontrivial due to the vague semantics of the *Unknown*, and this relaxation has little practical significance in an open-world where we have nothing about the prior knowledge of unknown actions. On the other hand, one may remove the *Unknown* or the *Background* class from training data, which are both infeasible under the OSTAL setting because (i) we have no temporal annotations of the unknown actions to remove them, and (ii) the pure background frames provide indispensable temporal context for action localization. Therefore, in contrast to the OSR problem, an unique technical challenge of OSTAL lies in distinguishing between actions of known and unknown classes, as well as the background frames.

Moreover, since the unknown actions are mixed with background frames without annotations, learning to distinguish foreground actions essentially reduces to a semi-supervised OSR problem [50, 67], that the model is trained with the labeled "known known" actions and the unlabeled "known unknown" actions while testing with data containing the "unknown unknown" actions[1].

To tackle these unique challenges, we propose to decouple the $(K+1)$-way action classification into K-way uncertainty-aware classification (Sec. 3.1) and actionness prediction (Sec. 3.2). Thus, we could address the first challenge above by jointly leveraging the uncertainty and actionness in a two-level decision-making (see Table 1) and the second challenge by the PU learning (Sec. 3.2).

---

[1]Refer to [16, 22] for more detailed discussions on these terminologies.

Table 1. **Our Motivations for the OSTAL.** The notations ↓ and ↑ denote small and large values, repsectively.

	Known Action	Unknown Action	Background
uncertainty (u)	↓	↑	↑
actionness (a)	↑	↑	↓

3.1. Action Classification

K**-way Uncertainty-aware Classification**   Following the existing Evidential Deep Learning (EDL) [4, 53], which is efficient to quantify the classification uncertainty, we assume a Dirichlet distribution $\text{Dir}(\mathbf{p}|\boldsymbol{\alpha})$ over the categorical probability $\mathbf{p} \in \mathbb{R}^K$, where $\boldsymbol{\alpha} \in \mathbb{R}^K$ is the Dirichlet strength. The EDL aims to directly predict $\boldsymbol{\alpha}$ by deep neural networks (DNNs). The model is trained by minimizing the following negative log-likelihood of data $\{x_i, y_i\}$:

$$\mathcal{L}_{\text{EDL}}^{(i)}(\boldsymbol{\alpha}_i) = \sum_{j=1}^{K} t_{ij}(\log(S_i) - \log(\alpha_{ij})), \quad (1)$$

where t_{ij} is a binary element of the one-hot form of label y_i, and $t_{ij} = 1$ only when $y_i = j$, and $S_i = \sum_j \alpha_{ij}$ is the total strength over K classes.

In testing, given the sample x_i^*, the action classification branch (DNN) produces non-negative evidence output $\mathbf{e}_i \in \mathbb{R}_+^K$. Then, the expectation of the classification probability is obtained by $\mathbb{E}[\mathbf{p}_i] = \boldsymbol{\alpha}_i / S_i$ where $\boldsymbol{\alpha}_i = \mathbf{e}_i + 1$ according to the evidence theory [54] and subjective logic [28]. And the classification uncertainty is estimated by $u_i = K / S_i$.

However, the above EDL method is empirically found ineffective in the OSTAL task since Eq. (1) gives equal consideration to each sample, which is practically not the case in OSTAL. In this paper, we propose to improve the generalization capability of EDL by encouraging the model to focus more on important samples in a principled way.

Momentum Importance-Balanced EDL   Inspired by the recent advances in imbalanced visual classification [32, 46], the sample importance can be measured by the influence function which is determined by the gradient norm. Specifically, let $\mathbf{h}_i \in \mathbb{R}^D$ be the feature input of the last DNN layer, an exponential evidence function is applied to predict the evidence, i.e., $\mathbf{e}_i \triangleq \exp(\mathbf{w}^T \mathbf{h}_i)$ where $\mathbf{w} \in \mathbb{R}^{D \times K}$ are the learnable weights of the DNN layer. The gradient $\mathbf{g}_i$ of the EDL loss $\mathcal{L}_{\text{EDL}}^{(i)}$ w.r.t. the logits $\mathbf{z}_i \triangleq \mathbf{w}^T \mathbf{h}_i$ is derived:

$$g_{ij} = \frac{\partial \mathcal{L}_{\text{EDL}}^{(ij)}}{\partial z_{ij}} = t_{ij}\left[\frac{S_i - K\alpha_{ij}}{S_i \alpha_{ij}}\right] = t_{ij}\left[\frac{1}{\alpha_{ij}} - u_i\right], \quad (2)$$

where the chain rule and the equality $u_i = K/S_i$ are used. Since $t_{ij} = 0$ when $j \neq y_i$, it is interesting to see a simple but meaningful gradient form, i.e., $g_{ik} = 1/\alpha_{ik} - u_i$ where $k = y_i$, and in our supplement we proved that $|g_{ik}| \in [0, 1)$.

Furthermore, inspired by [46], we consider the influence function given by the gradient norm of EDL loss w.r.t. the network parameters $\mathbf{w}$. According to the chain rule of $\mathbf{z}_i = \mathbf{w}^T \mathbf{h}_i$, the influence value ω_i can be derived:

$$\omega_i = \left(\sum_{k=1}^{K} |g_{ik}|\right)\left(\sum_{d=1}^{D} |h_{id}|\right) = \|\mathbf{g}_i\|_1 \cdot \|\mathbf{h}_i\|_1. \quad (3)$$

Detailed proof can be found in the supplement. We define the loss weight of sample x_i as the moving mean of influence values within the neighboring region of $\|\mathbf{g}_i\|_1$:

$$\tilde{\omega}_i^{(t)} = \epsilon \cdot \tilde{\omega}_i^{(t-1)} + (1 - \epsilon) \cdot \frac{1}{|\Omega_m|}\sum \Omega_m, \quad (4)$$

where Ω_m is a subset of ω_i whose gradient norm $\|\mathbf{g}_i\|_1$ falls into the m-th bin out of total M bins in the region $[0, 1]$, i.e., $\Omega_m = \{\omega_i | \|\mathbf{g}_i\|_1 \in [\frac{m-1}{M}, \frac{m}{M}], m = 1, \ldots, M\}$. The ϵ is a momentum factor within $[0, 1]$, M is a constant, and t is the training iteration. We set the initial weight $\tilde{\omega}_i^{(0)}$ as the 1.0. A larger ϵ means the set of influence values ω_i are less considered, while M controls the granularity of the neighborhood of the gradient norm. Eventually, the proposed Momentum Importance-Balanced (MIB) EDL loss is defined as:

$$\mathcal{L}_{\text{MIB-EDL}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{\omega}_i^{(t)} \mathcal{L}_{\text{EDL}}^{(i)}(\boldsymbol{\alpha}_i). \quad (5)$$

The proposed MIB-EDL loss encourages the model to smoothly focus on important samples as the training iteration increases. In practice, to stabilize the training, the re-weighting is applied after T_0 training iterations. Different to [46] that uses the inverse of ω_i to down-weight the influential samples for a balanced closed-set recognition, we use Eq. (3) to up-weight these samples for open-set recognition, and (4) to achieve a smooth update of the sample weight.

3.2. Actionness Prediction

Due to the mixture of unknown actions and pure background frames, it is not sufficient to distinguish between them by the evidential uncertainty over K known classes. Therefore, predicting the actionness that indicates how likely a sample is a foreground action is critical. We notice the fact that data from known classes are *positive* data while the samples from the "background" mixture are *unlabeled*. This intrinsically reduces to a semi-supervised learning problem called positive-unlabeled (PU) learning [5]. In this paper, we propose a simple yet effective PU learning method to predict the actionness.

Let $\hat{a}_i \in [0, 1]$ be the predicted actionness score of the sample x_i, the actionness in a training batch $\hat{\mathcal{A}} = \{\hat{a}_i\}$ can be splitted into the positive set $\hat{\mathcal{P}} = \{\hat{a}_i | y_i \geq 1\}$ and the unlabeled background set $\hat{\mathcal{U}} = \{\hat{a}_i | y_i = 0\}$. In

this paper, we propose to ascendingly sort the $\hat{\mathcal{U}}$ and select top-M samples to form the most likely negative set $\hat{\mathcal{N}} = \{\hat{a}_i | \hat{a}_i \in sort(\hat{\mathcal{U}})_{1,...,M}\}$. Then, a binary cross-entropy (BCE) loss could be applied to the $\hat{\mathcal{P}}$ and $\hat{\mathcal{N}}$:

$$\mathcal{L}_{\text{ACT}}(\hat{\mathcal{P}}, \hat{\mathcal{N}}) = -\frac{1}{|\hat{\mathcal{P}}|}\sum_{\hat{a}_i \in \hat{\mathcal{P}}} \log \hat{a}_i - \frac{1}{|\hat{\mathcal{N}}|}\sum_{\hat{a}_i \in \hat{\mathcal{N}}} \log(1 - \hat{a}_i). \tag{6}$$

Here, to achieve a balanced BCE training, we set the size of negative set to $M = |\hat{\mathcal{N}}| := \min(|\hat{\mathcal{P}}|, |\hat{\mathcal{U}}|)$ considering that in most training batches we have $|\hat{\mathcal{U}}| \gg |\hat{\mathcal{P}}|$. This BCE loss will push the probably pure background samples far away from positive actions. Though this method is straightforward, the learned actionness scores are found discriminative enough to distinguish between the foreground actions and background frames in the OSTAL setting (see Fig. 4a).

3.3. Location Regression

To maintain the flexibility of our method on existing TAL models, the temporal location regression follows the design of the TAL models. Take the state-of-the-art TAL model AFSD [36] as an example, it consists of a coarse stage to predict the location proposals $\hat{l}_i = [\hat{s}_i, \hat{e}_i]$ and a refined stage to predict the temporal offset $\hat{\delta}_i = [\hat{\delta}_i^{(s)}, \hat{\delta}_i^{(e)}]$ with respect to the $\hat{l}_i$. The coarse stage is learned by temporal Intersection-over-Union (tIoU) loss, while the refined stage is learned by an L_1 loss:

$$\begin{cases} \mathcal{L}_{\text{LOC}}(\{\hat{l}_i\}) = \frac{1}{N_C}\sum_i \mathbb{I}[y_i \geq 1]\left(1 - \frac{|\hat{l}_i \cap l_i|}{|\hat{l}_i \cup l_i|}\right) \\ \mathcal{L}_{\text{LOC}}(\{\hat{\delta}_i\}) = \frac{1}{N_R}\sum_i \mathbb{I}[y_i \geq 1](|\hat{\delta}_i - \delta_i|), \end{cases} \tag{7}$$

where N_C and N_R are corresponding number of samples that are matched with the ground truth action locations by an IoU threshold. The indicator function $\mathbb{I}[y_i \geq 1]$ filters out the unmatched samples which are treated as the "background" data. In testing, the predicted location is recovered by $l_i^* = [\hat{s}_i + 0.5(\hat{e}_i - \hat{s}_i)\hat{\delta}_i^{(s)}, \hat{e}_i + 0.5(\hat{e}_i - \hat{s}_i)\hat{\delta}_i^{(e)}]$. Note that our OpenTAL framework is not limited to specific TAL models but general in design.

3.4. IoU-aware Uncertainty Calibration

Though the loss functions defined by Eqs. (5)(6)(7) are sufficient for a complete OSTAL task, the learned uncertainty in the classification module is not calibrated by considering the localization performance. Intuitively, an action proposal of high temporal overlap with the ground truth location should contain more evidence and thus low uncertainty. To this end, we propose a novel IoU-aware uncertainty calibration method:

$$\mathcal{L}_{\text{IoUC}}^{(i)}(\hat{l}_i, u_i) = -w_{\hat{l}_i, l_i}\log(1 - u_i) - (1 - w_{\hat{l}_i, l_i})\log(u_i) \tag{8}$$

Algorithm 1 Inference Procedure
___
Require: Untrimmed test video.
Require: Trained OpenTAL model.
Require: Threshold τ from training data by Eq. (11)
 1: Data pre-processing (if applicable).
 2: Predict proposals $\mathcal{G} = \{l_i^*, \hat{y}_i, u_i, \hat{a}_i\}|_{i=1}^N$ by OpenTAL.
 3: Post-processing (if applicable).
 4: **for** each proposal $\mathcal{G}_i \in \mathcal{G}$ **do**
 5:     **if** $\hat{a}_i < 0.5$ **then**   ▷ Background
 6:         $\mathcal{G}_i$ is a *Background*; **continue**.
 7:     **end if**
 8:     **if** $u_i > \tau$ **then**   ▷ Unknown Action
 9:         $\mathcal{G}_i$ is *Unknown*.
10:     **else**   ▷ Known Action
11:         $\mathcal{G}_i$ is *Known* by $\hat{y}_i = \arg\max_j \mathbb{E}[\mathbf{p}_{ij}]$.
12:     **end if**
13: **end for**

---

where the weight w is a clipped form of the temporal IoU between the predicted and ground truth locations:

$$w_{\hat{l}_i, l_i} = \max\left(\gamma, \text{IoU}(\hat{l}_i, l_i)\right) \tag{9}$$

where the γ is a small non-negative constant. The cross-entropy form in Eq. (8) and (9) will encourage the model to produce high uncertainty ($u_i \to 1$) for action proposals with low localization quality ($w \to \gamma$).

The motivation behind the clipping by $\max()$ is that, given with the ground truth of known actions, both the proposals of background frames and unknown actions are not overlapped with the ground truth such that $\text{IoU}(\hat{l}_i, l_i) \leq 0$, the clipping could avoid reversing the loss value from positive to negative, while still maintaining a low localization quality γ. Besides, it is reasonable to encourage high uncertainty u_i by small γ for the location proposals of the background and the unknown actions in the OSTAL setting.

3.5. Training and Inference

The training procedure is to minimize the weighted sum of losses defined by Eqs. (5)(6)(7)(8):

$$\mathcal{L} = \mu\mathcal{L}_{\text{MIB-EDL}} + \mathcal{L}_{\text{ACT}} + \mathcal{L}_{\text{LOC}} + \mathbb{E}[\mathcal{L}_{\text{IoUC}}^{(i)}], \tag{10}$$

where μ is a hyperparameter, and $\mathbb{E}[\cdot]$ is to take the mean loss values over the input samples.

In inference, the untrimmed video input is fed into a TAL model, and our OpenTAL method trained on the TAL model could produce multiple action locations $\{l_i^*\}$, classification labels $\hat{y}_i = \arg\max_{j \in [1,...,K]} \mathbb{E}[\mathbf{p}_{ij}]$, classification uncertainty u_i, and actionness score $\hat{a}_i$. Together with the u_i and $\hat{a}_i$, a positively localized foreground action x_i, i.e., $a_i > 0.5$, can be accepted as known class $\hat{y}_i$, or rejected as

the unknown by the following simple scoring function:

$$P(x_i|a_i > 0.5) = \begin{cases} unknown, & \text{if } u_i > \tau, \\ \hat{y}_i, & \text{otherwise.} \end{cases} \quad (11)$$

The complete inference procedure is shown in Algorithm 1. In addition to this two-level decision, one-level decisions by the functional formulas of $P(x_i)$ w.r.t. to u_i and a_i are also plausible (see Table 5). However, we empirically found that Eq. (11) is the most effective formula while maintaining the explainable nature of decision-making.

4. Experiment

4.1. Implementation Details

Our method is implemented on the AFSD [36] model[2], which is a state-of-the-art TAL model. Pre-trained I3D [10] backbone is used in AFSD. The proposed OpenTAL is applied to both the coarse and refined stages of AFSD. Specifically, the proposed MIB re-weighting is applied after 10 training epochs. We empirically set the momentum ϵ to 0.99 and the number of bins M to 50. The small constant γ in Eq. (9) is set to 0.001. The loss weight μ in Eq. (10) is set to 10. We trained the model 25 epochs to ensure full convergence. The rest settings in AFSD are kept unchanged.

4.2. Datasets

THUMOS14 [27] and ActivityNet1.3 [8] are two commonly-used datasets for TAL evaluation. THUMOS14 dataset contains 200 training videos and 212 testing videos. ActivityNet1.3 dataset contains about 20K videos with 200 human activity categories. Since our method is not limited by data modality, we use RGB videos for training and testing by default. To enable OSTAL evaluation, we randomly select 3/4 THUMOS14 categories of the training videos as the known data. This random selection is repeated to generate three THUMOS14 open set splits between the known and the unknown. Considering that ActivityNet1.3 is newer and covers most THUMOS14 categories, the ActivityNet1.3 is not suitable to be the closed-set training data when the model is tested on THUMOS14. Therefore, we train models on the THUMOS14 known split, and use the THUMOS14 unknown split and the disjoint categories of ActivityNet1.3 as two sources open set testing data. To get the disjoint categories from ActivityNet1.3, we manually removed 14 semantically overlapping categories by referring to the THUMOS14 categories. Detailed dataset information could be found in our supplement.

4.3. Evaluation Protocols

The mean Average Precision (**mAP**) is typically used for the evaluation of closed set TAL performance. To enable OSTAL performance evaluation, the Area Under the

[2]https://github.com/TencentYoutuResearch/ActionDetection-AFSD

Receiver Operating Characteristic (**AUROC**) curve and the Area Under the Precision-Recall (**AUPR**) are introduced to evaluate the performance of detecting the unknown from the known actions for positively localized actions. To address the operational meaning in practice, we additionally report the False Alarm Rate at True Positive Rate of 95% (**FAR@95**), by which smaller value indicates better performance. However, we noticed the metrics above are insufficient for the OSTAL task because the multi-class classification performance of the known classes in the OSTAL setting is ignored. Inspired by the Open Set Classification Rate [13, 17, 43], we propose the Open Set Detection Rate (**OSDR**), which is defined as the area under the curve of Correct Detection Rate (CDR) and False Positive Rate (FPR). Given an operation point τ of the scoring function $P(x)$ for detecting the unknown and an operation point t_0 of the tIoU for localizing the foreground actions, CDR and FPR are defined as:

$$\begin{cases} \text{CDR}(\tau, t_0) = \dfrac{|\{x|(x \in \mathcal{F}_k) \wedge (\hat{f}_{x|y} = y) \wedge P(x) < \tau\}|}{|\mathcal{F}_k|} \\ \text{FPR}(\tau, t_0) = \dfrac{|\{x|(x \in \mathcal{F}_u) \wedge P(x) < \tau\}|}{|\mathcal{F}_u|} \end{cases} \quad (12)$$

where $\mathcal{F}_k$ is the set of positively localized known actions, i.e., $\mathcal{F}_k = \{x|(\text{tIoU} > t_0) \wedge (y \in [1, \ldots, K])\}$, and $\mathcal{F}_u$ is the set of positively localized unknown actions, i.e., $\mathcal{F}_u = \{x|(\text{tIoU} > t_0) \wedge (y = 0)\}$. The CDR indicates the fraction of known actions that are positively localized and correctly classified into their known classes, while the FPR denotes the fraction of unknown actions that are positively localized but falsely accepted as an arbitrary known class. Higher OSDR indicates better performance for the OSTAL task.

For stable evaluation, all results are reported by averaging the results of each evaluation metric over the three THUMOS14 splits. Results are reported at tIoU threshold 0.3 for THUMOS14 and 0.5 for ActivityNet1.4, and results by other thresholds are in the supplement.

4.4. Comparison with State-of-the-arts

The OpenTAL method is compared with the following baselines based on the AFSD: (1) **SoftMax**: use the softmax confidence score to identify the unknown. (2) **OpenMax**: use OpenMax [6] in testing to append the softmax scores with unknown class. (3) **EDL**: similar to [4], vanilla EDL method is used to replace the traditional cross-entropy loss for uncertainty quantification. Models are tested using both the THUMOS14 unknown spits and the ActivityNet1.3 disjoint subset. Results are reported in Table 2.

The results show that the OpenTAL outperforms the baselines by large margins on all OSTAL metrics, while still keeping comparable closed set TAL performance (less than 1% mAP decrease). The results also show that OpenMax

Table 2. **OSTAL Results (%).** Models trained on the THUMOS14 closed set are tested on the open sets by including the unknown classes from THUMOS14 and ActivityNet1.3, respectively. The mAP is provided as the reference of the TAL results on THUMOS14 closed set.

Methods	THUMOS14 as the Unknown				ActivityNet1.3 as the Unknown				mAP
	FAR@95 ($\downarrow$)	AUROC	AUPR	OSDR	FAR@95 ($\downarrow$)	AUROC	AUPR	OSDR	
SoftMax	85.58	54.70	31.85	23.40	85.05	56.97	53.54	27.63	55.81
OpenMax [6]	90.34	53.26	33.17	13.66	91.36	51.24	54.88	15.73	36.36
EDL [4]	81.42	64.05	40.05	36.26	84.01	62.82	53.97	38.56	52.24
OpenTAL	**70.96**	**78.33**	**58.62**	**42.91**	**63.11**	**82.97**	**80.41**	**50.49**	55.02

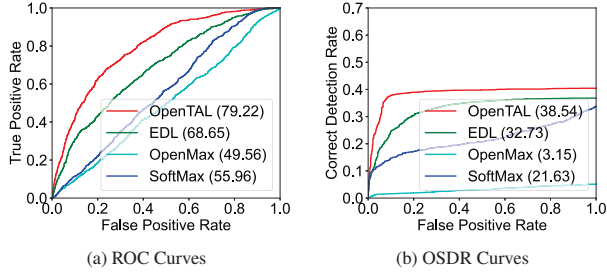(a) ROC Curves
(b) OSDR Curves

Figure 3. **ROC and OSDR curves on one THUMOS14 split.** Numbers in the brackets are AUROC or OSDR values.

does not work well on the OSTAL task, especially when the large-scale ActivityNet1.3 dataset is used as the unknown. The EDL works well but still far behind the proposed Open-TAL. Fig. 3 shows the detailed evaluation by the curves of AUROC and OSDR on one THUMOS14 split. Figures on other splits are in the supplement. They clearly show that the proposed OpenTAL on different operation points of scoring values and different open set splits is consistently better than the baselines.

4.5. Ablation Study

Component Ablation. By individually removing the major components of OpenTAL, three model variants are compared. (1) Without MIB: the proposed MIB re-weighting is removed so that the vanilla EDL loss (Eq. (1)) is used. (2) Without ACT: the actionness prediction is removed so that the $(K + 1)$-way classification in $\mathcal{L}_{\text{MIB-EDL}}$ (Eq. (5)) is adopted. (3) Without IoUC: the loss $\mathcal{L}_{\text{IoUC}}$ (Eq. (8)) is removed from the training. Results are reported in Table 3. They show that OpenTAL achieves the best performance. Specifically, the MIB re-weighting strategy contributes the most to the OSDR performance gain by around 30%. The actionness prediction (ACT) contributes the most to the FAR@95, AUROC, and AUPR metrics. Besides, the proposed IoUC loss also leads to significant performance gains on all metrics. These observations demonstrate the effectiveness of the three components for the OSTAL task.

Choices of Re-weighting Methods. We compare the proposed MIB re-weighting method (MIB (soft)) with the MIB (hard) and existing literature on sample re-weighting in Ta-

Table 3. **Ablation Results (%).** The proposed EDL re-weighting method (MIB), the actionness prediction (ACT), and the IoUC loss are individually ablated from the OpenTAL.

Variants	MIB	ACT	IoUC	FAR@95 ($\downarrow$)	AUROC	AUPR	OSDR
(1)		✓	✓	77.20	76.41	56.65	12.10
(2)	✓		✓	82.85	58.12	31.80	37.89
(3)	✓	✓		79.64	62.73	37.86	39.39
OpenTAL	✓	✓	✓	**70.96**	**78.33**	**58.62**	**42.91**

Table 4. **Results of Different Re-weightings (%).** MIB (hard) means the momentum factor $\epsilon = 0$ in Eq. (4) such that the sample weight is updated in a hard manner, while the MIB (soft) sets the ϵ to 0.99 to enable a soft update, and wo. Re-weight means $\epsilon = 1.0$.

Methods	FAR@95 ($\downarrow$)	AUROC	AUPR	OSDR
wo. Re-weight.	77.20	76.41	56.65	12.10
Focal [39]	91.05	56.67	35.55	2.04
GHM [35]	78.33	73.52	54.03	1.41
IB [46]	80.23	75.91	58.00	2.18
MIB (hard)	**66.34**	78.16	57.66	38.90
MIB (soft)	70.96	**78.33**	**58.62**	**42.91**

ble 4. The results show that the focal loss (Focal) [39] does not work well with the OpenTAL framework. GHM [35] and IB [46] methods could achieve comparable FAR@95, AUROC, and AUPR performance, but their OSDR results are still largely far behind ours. Note that these methods are all designed for closed set recognition, thus the proposed MIB is more suitable for open set scenarios. Besides, the hard version of MIB that the momentum mechanism is removed by setting the ϵ to 0, could improve about 4% FAR@95 while sacrifice the AUROC, AUPR, and OSDR.

Choices of Scoring Function. The scoring function is critical to identify the known and unknown actions, as well as the background frames in model inference. In addition to the proposed two-level decision by (11), we compare it with four reasonable one-level decision methods by utilizing actionness a_i and uncertainty u_i. The results in Table 5 show that using the maximum classification confidence (the 1st row) or other compositions of u_i and a_i (the 2nd and 3rd rows) cannot achieve favorable performance.

Table 5. **Scoring Functions.** It shows when conditioned on $a_i >$ 0.5, uncertainty u_i is the best scoring function for the OSTAL task.

Scoring Functions	FAR@95 ($\downarrow$)	AUROC	AUPR	OSDR
$P(x_i) = 1 - \max_j(\boldsymbol{\alpha}_i/S_i)$	77.90	59.50	35.82	31.38
$P(x_i) = u_i/(1 - a_i)$	79.16	61.94	38.52	30.64
$P(x_i) = a_i/(1 - u_i)$	90.39	72.71	56.19	38.24
$P(x_i) = u_i \cdot a_i$	**70.64**	77.52	58.17	42.44
$P(x_i\|a_i > 0.5) = u_i$	70.96	**78.33**	**58.62**	**42.91**

The proposed method (the last row) is slightly better than the product between u_i and a_i (the 4-th row) with comparable FAR@95 performance. Though there are certainly other alternatives, our scoring function achieves the best performance while maintaining a good decision-making explanation, which means that the foreground actions are identified first by a_i, based on which the known and unknown actions are further distinguished by u_i.

Distributions of Actionness and Uncertainty. To show the quality of the learned actionness and uncertainty, we visualized their distributions on the test set in Fig. 4. Specifically, the dominant modes in Fig. 4a show that foreground actions are majorly assigned with high actionness while the background frames are with low actionness, and the dominant modes in Fig 4b show that the actions of known classes are majorly assigned with low uncertainty while those of the unknowns are with high uncertainty. These observations align well with the expectation of our OpenTAL method.

Qualitative Results. Fig. 5 shows the qualitative results of the proposed OpenTAL and baseline approaches. The three video samples are from the THUMOS14 dataset. The results clearly show that OpenTAL is superior to baselines in terms of both recognizing the known actions (colored segments in the 1st video), and rejecting the unknown actions (black segments in the 2nd and 3rd videos).

Limitations. We note that all those methods are not showing remarkable high OSDR performance, which indicates the challenging nature of the OSTAL task and there exists large room for improvement in the OpenTAL.

5. Conclusion

In this paper, we introduce the Open Set Temporal Action Localization (OSTAL) task. It aims to simultaneously localize and recognize human actions, and to reject the unknown actions from untrimmed videos in an open-world. The unique challenge lies in discriminating between known and unknown actions as well as background video frames. To this end, we propose a general OpenTAL framework to enable existing TAL models for the OSTAL task. The OpenTAL predicts the locations, classifications with uncertainties, and actionness to jointly achieve the goal. For com-

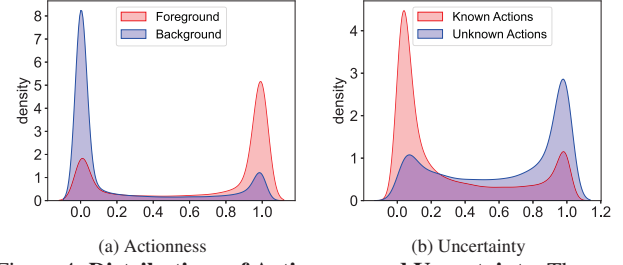(a) Actionness   (b) Uncertainty

Figure 4. **Distributions of Actionness and Uncertainty.** The two figures show significant separation between the foreground actions and background frames by actionness score, as well as the separation between the known and unknown actions by uncertainty.
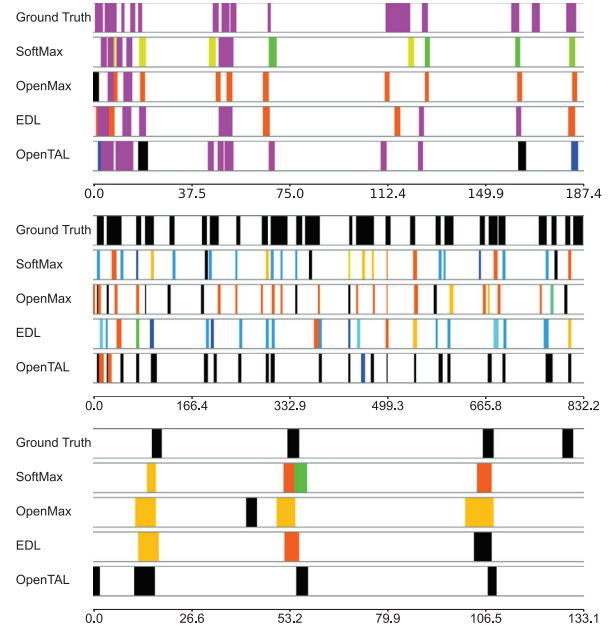

Figure 5. **Qualitative Results.** We show the actions of unknown classes with black color, while the rest colors are actions of known classes. The x-axis represents the timestamps (seconds).

prehensive OSTAL evaluation, the Open Set Detection Rate is introduced. The OpenTAL is empirically demonstrated to be effective and significantly outperform existing baselines. We believe the generality of the OpenTAL design could inspire relevant research fields such as spatio-temporal action detection, video object detection, and video grounding toward open set scenarios.

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *NeurIPS*, 2020. 2

[2] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *ECCV*, 2020. 2

[3] Wentao Bao, Qi Yu, and Yu Kong. Object-aware centroid voting for monocular 3d object detection. In *IROS*, 2020. 1

[4] Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. In *ICCV*, 2021. 1, 2, 4, 6, 7

[5] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020. 2, 4

[6] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *CVPR*, 2016. 2, 6, 7

[7] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *BMVC*, 2017. 2

[8] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2, 6

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1

[10] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017. 1, 6

[11] Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Deep metric learning for open world semantic segmentation. In *ICCV*, 2021. 2

[12] 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. 1

[13] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2021. 2, 6

[14] 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*, 2020. 2

[15] Junwen Chen, Wentao Bao, and Yu Kong. Group activity prediction with sequential relational anticipation model. In *ECCV*, 2020. 1

[16] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, 2020. 2, 3

[17] Akshay Raj Dhamija, Manuel Günther, and Terrance E Boult. Reducing network agnostophobia. In *NeurIPS*, 2018. 6

[18] Luke Ditria, Benjamin J Meyer, and Tom Drummond. Open-GAN: Open set generative adversarial networks. In *ACCV*, 2020. 2

[19] Zhen Fang, Jie Lu, Anjin Liu, Feng Liu, and Guangquan Zhang. Learning bounds for open-set learning. In *ICML*, 2021. 2

[20] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1

[21] Zongyuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative OpenMax for multi-class open set classification. In *BMVC*, 2017. 2

[22] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE TPAMI*, 2020. 3

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[24] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *CVPR*, 2016. 2

[25] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *CVPR*, 2021. 2

[26] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, 2014. 2

[27] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014. 2, 6

[28] Audun Jøsang. *Subjective logic*. Springer, 2016. 4

[29] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 2

[30] Pedro Ribeiro Mendes Júnior, Terrance E Boult, Jacques Wainer, and Anderson Rocha. Specialized support vector machines for open-set recognition. *arXiv preprint arXiv:1606.03802*, 2016. 2

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[32] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017. 4

[33] Shu Kong and Deva Ramanan. OpenGAN: Open-set recognition via open data generation. In *ICCV*, 2021. 2

[34] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018. 1

[35] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *AAAI*, 2019. 7

[36] 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. 1, 2, 3, 5, 6

[37] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, 2019. 2

[38] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, 2018. 2

[39] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 7

[40] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, 2019. 2

[41] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, 2018. 2

[42] Martin Mundt, Iuliia Pliushch, Sagnik Majumder, and Visvanathan Ramesh. Open set recognition through deep neural network uncertainty: Does out-of-distribution detection require generative classifiers? In *ICCVW*, 2019. 2

[43] Lawrence Neal, Matthew Olson, Xiaoli Fern, Weng-Keen Wong, and Fuxin Li. Open set learning with counterfactual images. In *ECCV*, 2018. 6

[44] Hugo Oliveira, Caio Silva, Gabriel LS Machado, Keiller Nogueira, and Jefersson A dos Santos. Fully convolutional open set segmentation. *Machine Learning*, pages 1–52, 2021. 2

[45] Poojan Oza and Vishal M Patel. C2AE: Class conditioned auto-encoder for open-set recognition. In *CVPR*, 2019. 2

[46] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, 2021. 4, 7

[47] Pramuditha Perera, Vlad I Morariu, Rajiv Jain, Varun Manjunatha, Curtis Wigington, Vicente Ordonez, and Vishal M Patel. Generative-discriminative feature representations for open-set recognition. In *CVPR*, 2020. 2

[48] Trung Pham, Thanh-Toan Do, Gustavo Carneiro, Ian Reid, et al. Bayesian semantic instance segmentation in open set world. In *ECCV*, 2018. 2

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2017. 1

[50] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. In *NeurIPS*, 2021. 3

[51] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE TPAMI*, 35(7):1757–1772, 2012. 1, 2

[52] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE TPAMI*, 36(11):2317–2324, 2014. 2

[53] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, 2018. 2, 4

[54] Kari Sentz, Scott Ferson, et al. *Combination of evidence in Dempster-Shafer theory*, volume 4015. Sandia National Laboratories Albuquerque, 2002. 4

[55] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, 2017. 2

[56] Yu Shu, Yemin Shi, Yaowei Wang, Tiejun Huang, and Yonghong Tian. p-odn: prototype-based open deep network for open set recognition. *Scientific reports*, 10(1):1–13, 2020. 2

[57] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. Class semantics-based attention for action detection. In *ICCV*, 2021. 2

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

[59] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *ICCV*, 2021. 2

[60] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. In *ICCV*, 2021. 2

[61] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 2

[62] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, 2020. 1, 2

[63] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. In *CVPR*, 2018. 2

[64] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, Qing Yang, and Cheng-Lin Liu. Convolutional prototype network for open set recognition. *IEEE TPAMI*, 2020. 2

[65] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016. 2

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

[67] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *ECCV*, 2020. 3

[68] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *CVPR*, 2021. 2

[69] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, 2019. 1, 2

[70] Hui Zhang and Henghui Ding. Prototypical matching and open set rejection for zero-shot semantic segmentation. In *ICCV*, 2021. 2

[71] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, 2020. 2

[72] Xujiang Zhao, Yuzhe Ou, Lance Kaplan, Feng Chen, and Jin-Hee Cho. Quantifying classification uncertainty using regularized evidential neural networks. In *AAAI*, 2019. 2

[73] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 2

[74] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Learning placeholders for open-set recognition. In *CVPR*, 2021. 2

[75] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *ICCV*, 2021. 2

[76] Yixiong Zou, Shanghang Zhang, Guangyao Chen, Yonghong Tian, Kurt Keutzer, and José M. F. Moura. Annotation-efficient untrimmed video action recognition. In *ACM MM*, 2021. 2