JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

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

**Mengyuan Chen, Junyu Gao, Changsheng Xu**

State Key Laboratory of Multimodal Artificial Intelligence Systems
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Peng Cheng Laboratory

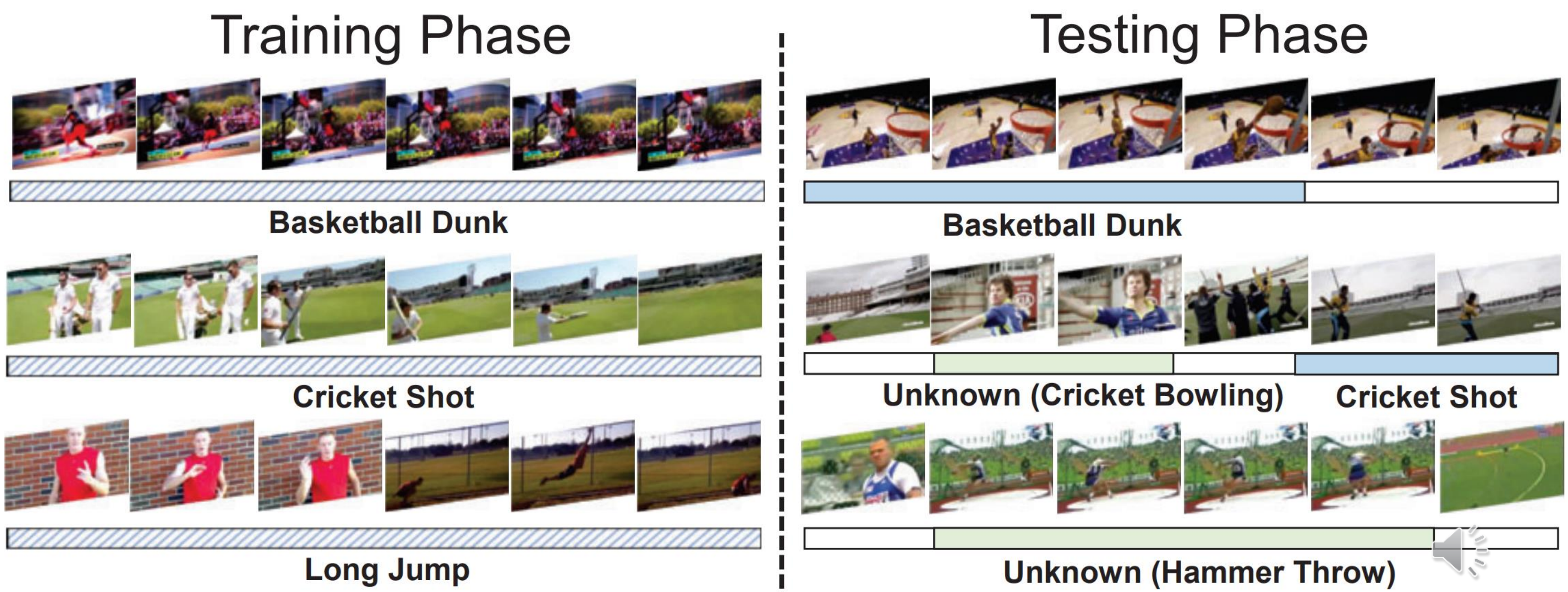
2023-06

WED-PM-228

# Introduction

## ■ Open-world Weakly-supervised TAL (OWTAL)

- ➢ With only video-level labels for training, OWTAL aims to localize both known and unknown action instances in testing videos.

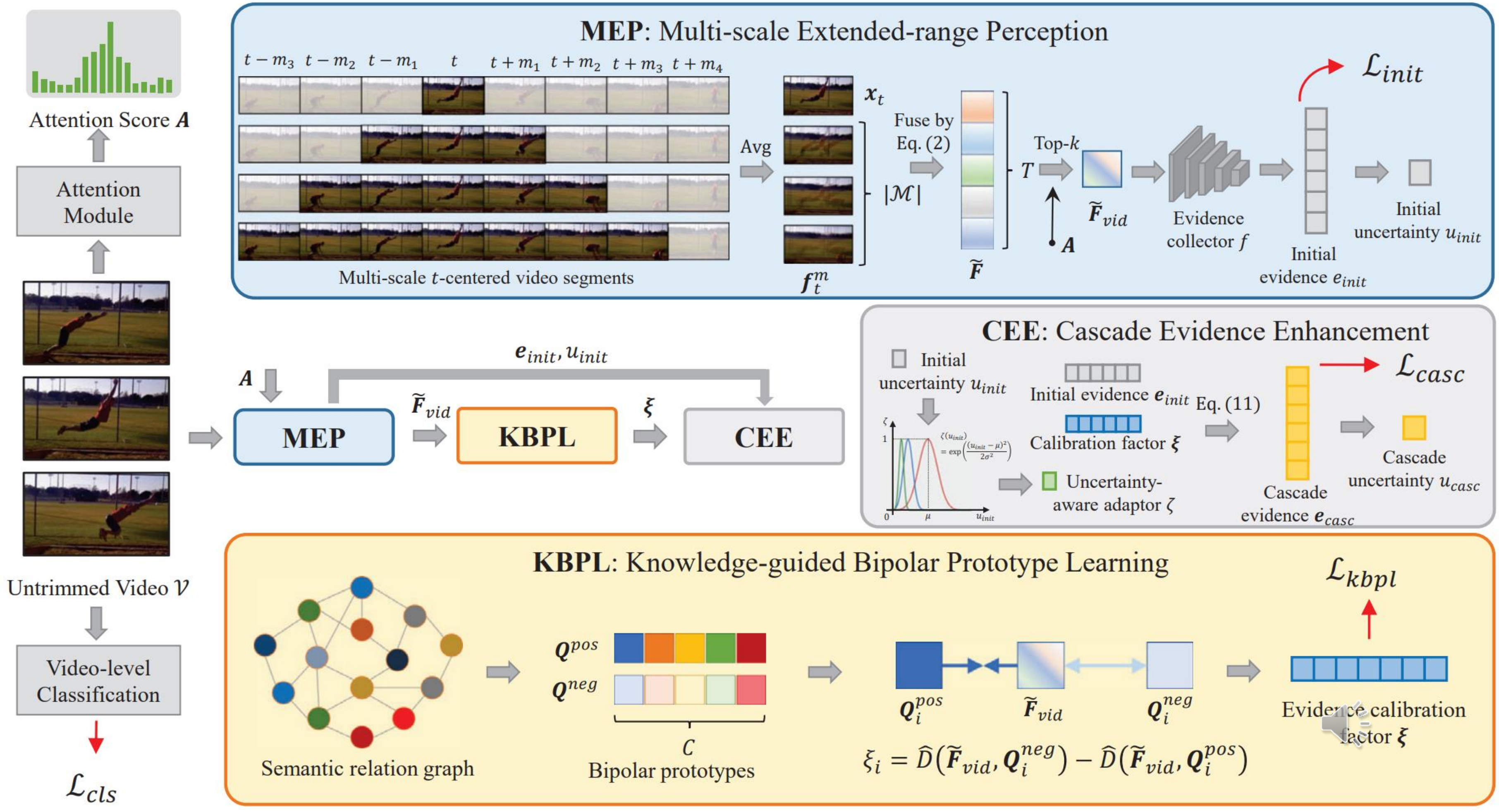


# Challenges

- **Ambiguity of annotations of closed-set (known) action instances**
  - Previous works indicate that the closed-set and open-set performance are highly correlated. 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.
- **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.

# Method

## ■ Cascade Evidential Learning



# Method

## ■ Multi-scale Extended-range Perception for Initial Evidence Collection

### ➢ Multi-scale Extended-range Perception

$$\boldsymbol{f}_t^m = \frac{1}{e_t^m - s_t^m + 1} \sum_{s_t^m \leq i \leq e_t^m} \boldsymbol{x}_i. \qquad \tilde{\boldsymbol{f}}_t = (1-\alpha_t)\varphi_1(\boldsymbol{x}_t) + \alpha_t \sum_{m \in \mathcal{M}} \delta(\omega_t^m)\varphi_2(\boldsymbol{f}_t^m)$$

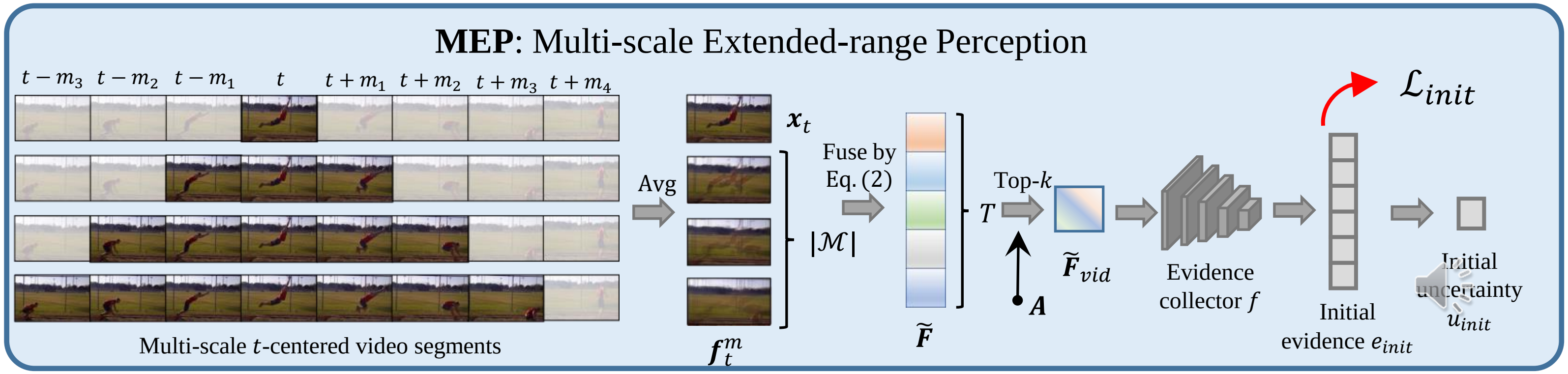
$$\alpha_t = \frac{1}{2|\mathcal{M}|} \sum_{m \in \mathcal{M}} (\omega_t^m + 1)\,, \omega_t^m = \cos\left(\frac{\boldsymbol{x}_t}{|\boldsymbol{x}_t|}, \frac{\boldsymbol{f}_t^m}{|\boldsymbol{f}_t^m|}\right)$$

### ➢ Initial Evidence Collection

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

### ➢ Loss function for evidential learning

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



# Method

## ■ Knowledge-guided Bipolar Prototype Learning for Evidence Calibration Factors

- **Graph network update**

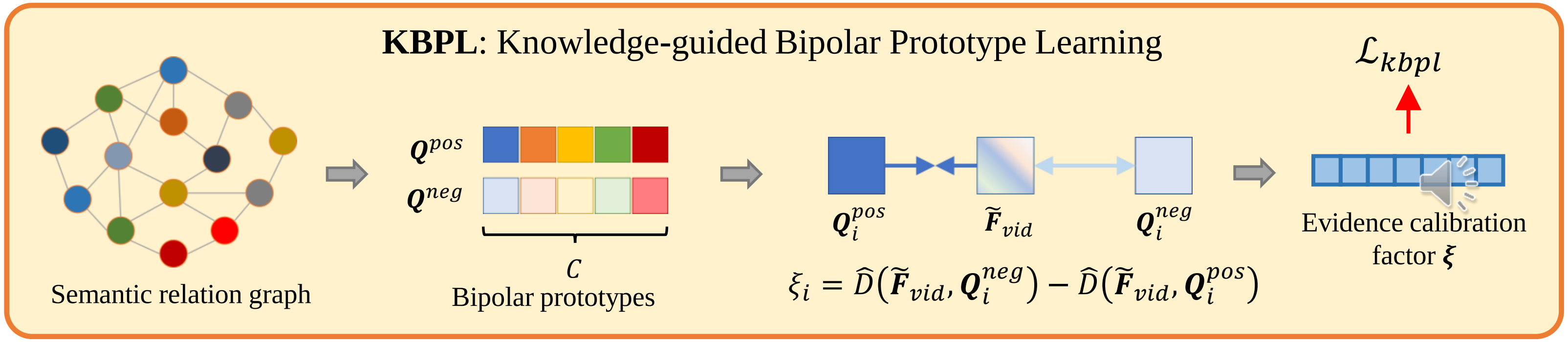
$$\tilde{\boldsymbol{n}}_j = h_2\Big(\underset{z\in\{1,2,3\}}{\text{Concat}}\Big(\sum_{l\in\mathcal{S}_{j,z}} \delta(\eta_{j,l}) h_1(\boldsymbol{n}_l)\Big)\Big) + h_3(\boldsymbol{n}_j)$$

- **Distance to bipolar prototypes**

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

- **Loss function**

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



# Method

## ■ Cascade Evidence Enhancement

- **Uncertainty-aware adaptor**

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

- **Evidence calibration**

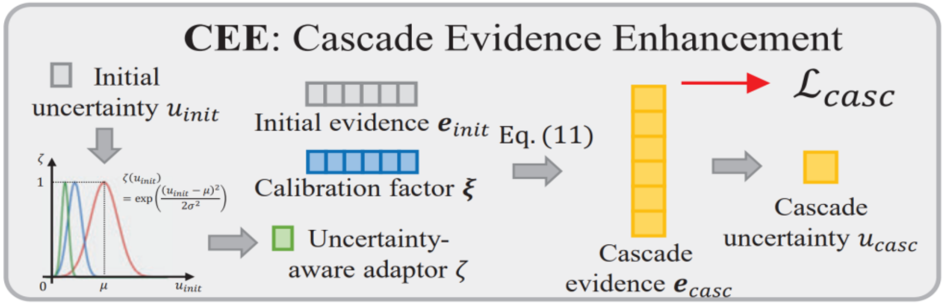
$$e_{casc,i}=(1+\zeta(u_{init})\tanh(\xi_i))e_{init,i}$$

- **Total loss function**

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

$$where\ \alpha_{casc,i}=e_{casc,i}+1, S_{casc}=\sum_{i}\alpha_{casc,i}$$

$$\mathcal{L}=\mathcal{L}_{cls}+\mathcal{L}_{init}+\mathcal{L}_{kbpl}+\mathcal{L}_{casc}$$



# Method

## ■ Inference

**Algorithm 1** Action Categories Inference Procedure

**Input**: Untrimmed testing video $\mathcal{V}$.
**Require**: Trained CELL model.
**Require**: Threshold $\tau$ obtained from training data (Please refer to **Implementation Details**).
**Output**: Set $\Psi$ 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** $\Psi$

# Experiments

- **Comparison results on THUMOS14**

| 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** |

# Experiments

## Comparison results on ActivityNet-v1.3

| 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** |

## Ablation study

| 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** |

# Thanks!


Any problem, please feel free contact me:

Mengyuan Chen
chenmengyuan2021@ia.ac.cn