OpenTAL: Towards Open Set Temporal Action Localization

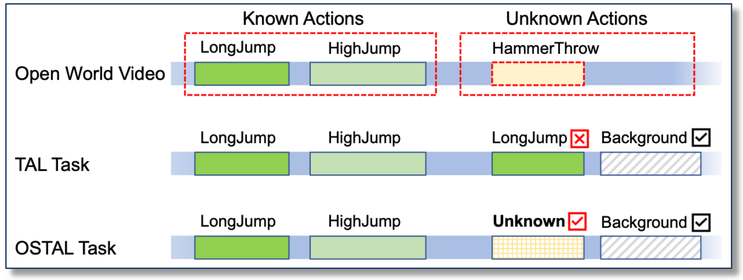
Wentao Bao, Qi Yu, Yu Kong

Golisano College of Computing and Information Sciences, Rochester Institute of Technology (RIT)

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Problem Setup

- Open Set Temporal Action Localization (**OSTAL**) aims to:
  - **Localize** human actions temporally in untrimmed videos.
  - **Recognize** the known types of the actions.
  - **Reject** the unknown actions.
- Model Training & Testing
  - Trained with known classes (**closed-set**)
  - Tested with known & unknown classes (**open-set**)
- Comparison with TAL



Challenges

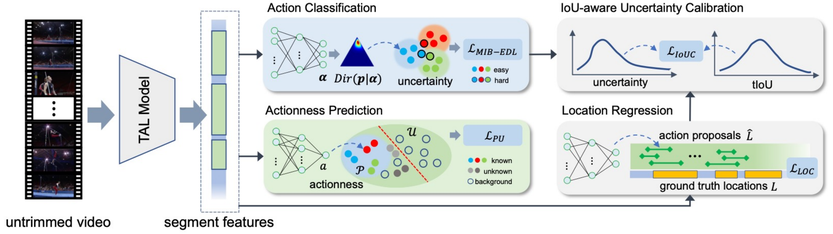
- ***Background*** CANNOT be removed (provide context for localization).
- ***Unknown*** actions CANNOT be removed (no temporal annotations).
- ***Semi-supervision:*** unknown actions are MIXED with backgrounds.

Significance

- The first attempt for temporal action localization in open-set setting.
- Valuable task for real-world landing of video TAL models:
  - Localize unexpected new video events over time.
  - Facilitate large-scale temporal annotations.

OpenTAL Model

- Overview of Framework



- **Learning:** decoupling the OSTAL objective into:
  - Uncertainty-aware Action Classification: **"Are they known / unknown?"**

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

  - Actionness Prediction: **"Are they foreground / background?"**

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

  - Temporal Location Regression: **"Where are the human actions?"**

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

  - IoU-aware Uncertainty Calibration: **"Is the uncertainty consistent with tIoU?"**

$$\mathcal{L}_{\text{IoUC}}^{(i)}(\hat{l}_i, u_i) = -w_{\hat{l}_i,l_i}\log(1-u_i) - (1-w_{\hat{l}_i,l_i})\log(u_i)$$
$$w_{\hat{l}_i,l_i} = \max\left(\gamma, \text{IoU}(\hat{l}_i, l_i)\right)$$

- **Inference**: jointly using **uncertainty** $u = K/\sum_k \hat{\alpha}_k$ and **actionness** $\hat{a}$.

Experimental Results

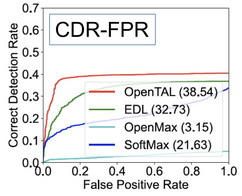
- **New Metric**: Open Set Detection Rate (**OSDR**)

Area Under the CDR-FPR curve

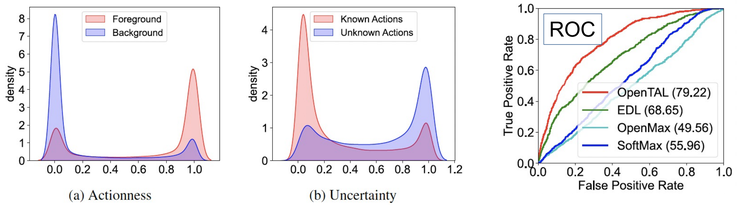
$$\begin{cases}\text{CDR}(\tau, t_0) = \dfrac{|\{x|(x\in\mathcal{F}_k)\wedge(\hat{f}_{x|y}=y)\wedge P(x)<\tau\}|}{|\mathcal{F}_k|} \\ \text{FPR}(\tau, t_0) = \dfrac{|\{x|(x\in\mathcal{F}_u)\wedge P(x)<\tau\}|}{|\mathcal{F}_u|}\end{cases}$$

- **Compare with SOTA**

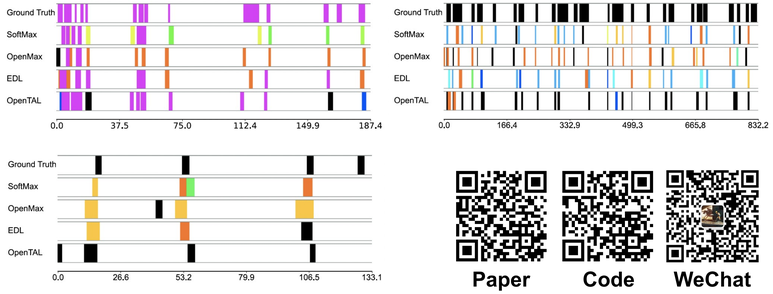
Methods	THUMOS14 as the Unknown			
	FAR@95 (↓)	AUROC	AUPR	OSDR
SoftMax	85.58	54.70	31.85	23.40
OpenMax [6]	90.34	53.26	33.17	13.66
EDL [4]	81.42	64.05	40.05	36.26
OpenTAL	**70.96**	**78.33**	**58.62**	**42.91**



- **Distributions of Actionness and Uncertainty**



Demo



Paper Code WeChat