Weakly-supervised Temporal Action Localization by Uncertainty Modeling

Pilhyeon Lee[1] Jinglu Wang[2] Yan Lu[2] Hyeran Byun[1,3*]
[1]Department of Computer Science, Yonsei University
[2]Microsoft Research Asia [3]Graduate School of AI, Yonsei University

GitHub

Weakly-supervised temporal action localization

❑ **Goal**
- predicting action intervals and their categories in videos

❑ **Given**
- video-level labels
(about what action categories an input video contains)

❑ **Challenge**
- how to model background in the weakly-supervised setting

Proposed Method

❑ **Motivation**
- We observe that background frames are largely inconsistent and unconstrained.
- Based on the observation, we propose to consider them **out-of-distribution** samples ($d = 0$).
(Accordingly, actions are considered **in-distribution** ($d = 1$).)

❑ **Uncertainty modeling**
- We add the out-of-distribution detection term to the conventional action classification term at segment-level.

$$P(y_{n,t} = c|\tilde{s}_{n,t}) = P(y_{n,t} = c, d = 1|\tilde{s}_{n,t}) = \underbrace{P(y_{n,t} = c|d = 1, \tilde{s}_{n,t})}_{\text{In-distribution classification}} \underbrace{P(d = 1|\tilde{s}_{n,t})}_{\text{Out-of-distribution detection}}$$

- The second term is estimated by the feature magnitude.

$$P(d = 1|\tilde{s}_{n,t}) = \frac{\min(m, \|f_{n,t}\|)}{m},$$

where m is the pre-defined maximum magnitude.

❑ **Training objectives** $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{um}} + \beta\mathcal{L}_{\text{be}}$
- Video-level loss

$$: \mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} -y_{n;c}\log p_c(v_n)$$

- Uncertainty modeling loss

$$: \mathcal{L}_{\text{um}} = \frac{1}{N}\sum_{n=1}^{N}(\max(0, m - \|f_n^{\text{act}}\|) + \|f_n^{\text{bkg}}\|)^2$$

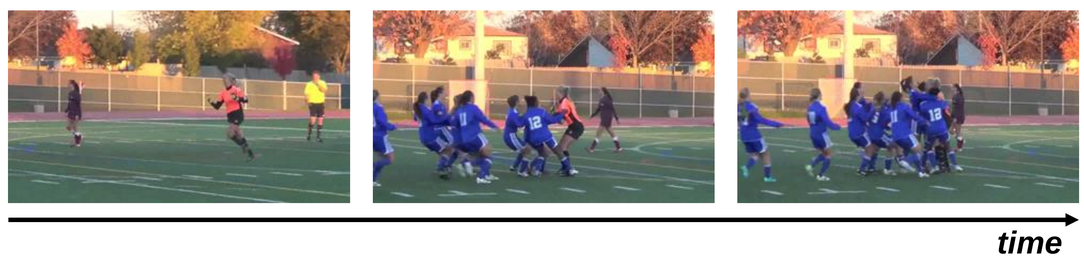
- Background entropy loss

$$: \mathcal{L}_{\text{be}} = \frac{1}{NC}\sum_{n=1}^{N}\sum_{c=1}^{C} -\log(p_c(\tilde{s}_n^{\text{bkg}}))$$

Existing approaches to background modeling

❑ **Staticity assumption**
- It is assumed that all background frames are static.
- Static frames are concatenated to generate pseudo background videos for training.

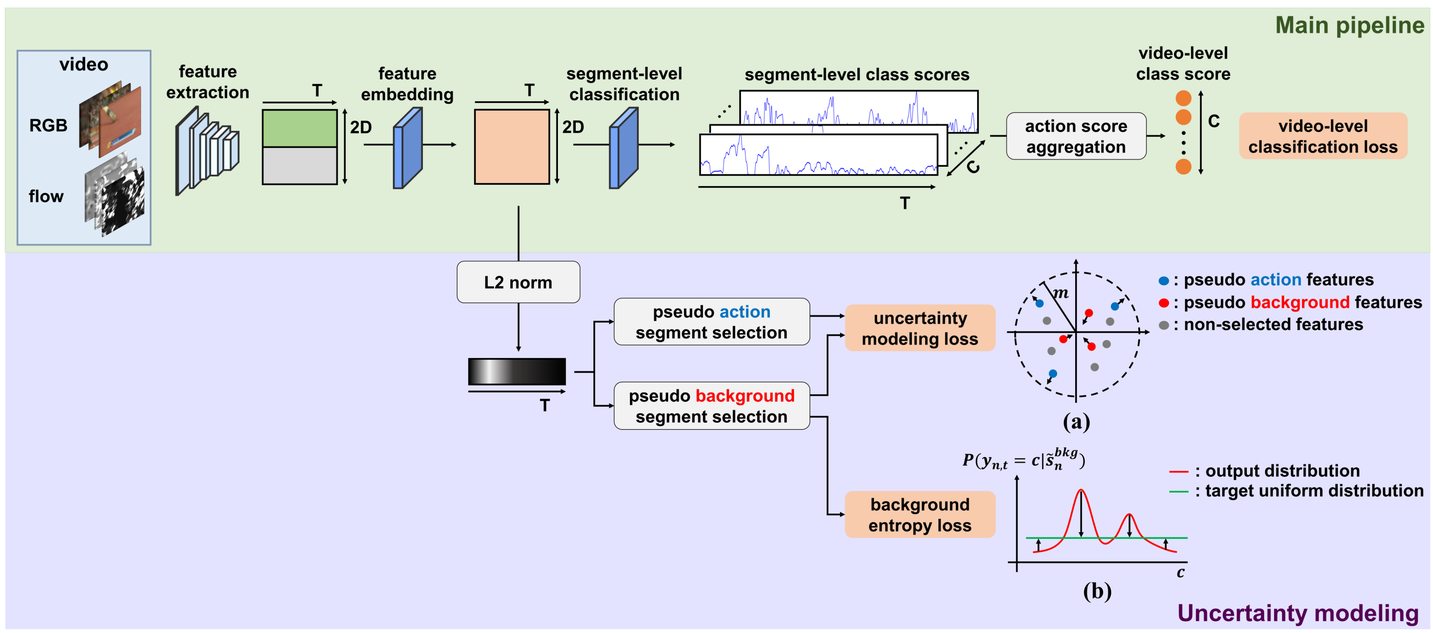
➔ The assumption does not necessarily hold true, as background frames may be dynamic (as shown in the figure below).



❑ **Auxiliary class**
- Background frames are classified as the $(C+1)$-*th* class, where C is the number of action classes.

➔ It is infeasible to push all of them to a single class, as background frames have no common semantics (see the figure below).

❑ **Overall architecture**



Experimental Results

❑ **Comparison with state-of-the-art methods**

Supervision	Method	mAP@IoU (%) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
Full	S-CNN (Shou, Wang, and Chang 2016)	47.7	43.5	36.3	28.7	19.0	10.3	5.3	27.3
	SSN (Zhao et al. 2017)	66.0	59.4	51.9	41.0	29.8	-	-	-
	TAL-Net (Chao et al. 2018)	59.8	57.1	53.2	48.5	42.8	33.8	20.8	45.1
	BSN (Lin et al. 2018)	-	-	53.5	45.0	36.9	28.4	20.0	-
	P-GCN (Zeng et al. 2019)	69.5	67.8	63.6	57.8	49.1	-	-	-
	G-TAD (Xu et al. 2020)	-	-	66.4	60.4	51.6	37.6	22.9	-
Weak†	STAR (Xu et al. 2019)	68.8	60.0	48.7	34.7	23.0	-	-	-
	3C-Net (Narayan et al. 2019)	59.1	53.5	44.2	34.1	26.6	-	8.1	-
	PreTrimNet (Zhang et al. 2020)	57.5	50.7	41.4	32.1	23.1	14.2	7.7	23.7
Weak	UntrimmedNets (Wang et al. 2017)	44.4	37.7	28.2	21.1	13.7	-	-	-
	Hide-and-seek (Singh and Lee 2017)	36.4	27.8	19.5	12.7	6.8	-	-	-
	STPN (Nguyen et al. 2018)	52.0	44.7	35.5	25.8	16.9	9.9	4.3	27.0
	AutoLoc (Shou et al. 2018)	-	-	35.8	29.0	21.2	13.4	5.8	-
	W-TALC (Paul, Roy, and Roy-Chowdhury 2018)	55.2	49.6	40.1	31.1	22.8	-	7.6	-
	MAAN (Yuan et al. 2019)	59.8	50.8	41.1	30.6	20.3	12.0	6.9	31.6
	Liu *et al.* (Liu, Jiang, and Wang 2019)	57.4	50.8	41.2	32.1	23.1	15.0	7.0	32.4
	CleanNet (Liu et al. 2019)	-	-	37.0	30.9	23.9	13.9	7.1	-
	TSM (Yu et al. 2019)	-	-	39.5	-	24.5	-	7.1	-
	Nguyen *et al.* (Nguyen, Ramanan, and Fowlkes 2019)	60.4	56.0	46.6	37.5	26.8	17.6	9.0	36.3
	BaS-Net (Lee, Uh, and Byun 2020)	58.2	52.3	44.6	36.0	27.0	18.6	10.4	35.3
	RPN (Huang et al. 2020)	62.3	57.0	48.2	37.2	27.9	16.7	8.1	36.8
	DGAM (Shi et al. 2020)	60.0	54.2	46.8	38.2	28.8	19.8	11.4	37.0
	Gong *et al.* (Gong et al. 2020)	-	-	46.9	38.9	30.1	19.8	10.4	-
	ActionBytes (Jain, Ghodrati, and Snoek 2020)	-	-	43.0	35.8	29.0	-	9.5	-
	EM-MIL (Luo et al. 2020)	59.1	52.7	45.5	36.8	30.5	22.7	**16.4**	37.7
	A2CL-PT (Min and Corso 2020)	61.2	56.1	48.1	39.0	30.1	19.2	10.6	37.8
	TSCN (Zhai et al. 2020)	63.4	57.6	47.8	37.7	28.7	19.4	10.2	37.8
	Ours	**67.5**	**61.2**	**52.3**	**43.4**	**33.7**	**22.9**	12.1	**41.9**

THUMOS'14

Sup.	Method	mAP@IoU (%) 0.5	0.75	0.95	AVG
Full	SSN (2017)	41.3	27.0	6.1	26.6
Weak†	3C-Net (2019)	37.2	-	-	21.7
Weak	UntrimmedNets (2017)	7.4	3.2	0.7	3.6
	AutoLoc (2018)	27.3	15.1	3.3	16.0
	W-TALC (2018)	37.0	12.7	1.5	18.0
	Liu *et al.* (2019)	36.8	22.9	5.6	22.4
	CleanNet (2019)	37.1	20.3	5.0	21.6
	TSM (2019)	28.3	17.0	3.5	17.1
	RPN (2020)	37.6	23.9	5.4	23.3
	BaS-Net (2020)	38.5	24.2	5.6	24.3
	DGAM (2020)	41.0	23.5	5.3	24.4
	Gong *et al.* (2020)	40.0	25.0	4.6	24.6
	EM-MIL (2020)	37.4	-	-	20.3
	TSCN (2020)	37.6	23.7	5.7	23.6
	Ours	**41.2**	**25.6**	**6.0**	**25.9**

ActivityNet1.2

Sup.	Method	mAP@IoU (%) 0.5	0.75	0.95	AVG
Full	TAL-Net (2018)	38.2	18.3	1.3	20.2
	BSN (2018)	46.5	30.0	8.0	30.0
	BMN (2019)	50.1	34.8	8.3	33.9
	P-GCN (2019)	48.3	33.2	3.3	31.1
	G-TAD (2020)	50.4	34.6	9.0	34.1
Weak†	STAR (2019)	31.1	18.8	4.7	-
	PreTrimNet (2020)	34.8	20.9	5.3	22.5
Weak	STPN (2018)	29.3	16.9	2.6	-
	MAAN (2019)	33.7	21.9	5.5	-
	Liu *et al.* (2019)	34.0	20.9	**5.7**	21.2
	TSM (2019)	30.3	19.0	4.5	-
	Nguyen *et al.* (2019)	36.4	19.2	2.9	-
	BaS-Net (2020)	34.5	22.5	4.9	22.2
	A2CL-PT (2020)	36.8	22.5	5.2	22.5
	TSCN (2020)	35.3	21.4	5.3	21.7
	Ours	**37.0**	**23.9**	**5.7**	**23.7**

ActivityNet1.3

❑ **Ablation study**

Score		Loss		mAP@IoU (%) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG
softmax	fusion	$\mathcal{L}_{\text{um}}$	$\mathcal{L}_{\text{be}}$								
✓				42.3	35.1	27.7	21.7	16.4	10.5	4.6	22.6
	✓			58.8	50.0	40.5	31.7	23.6	15.5	7.2	32.5
	✓	✓		65.3	59.2	50.8	42.3	32.4	21.1	10.4	40.2
	✓	✓	✓	67.5	61.2	52.3	43.4	33.7	22.9	12.1	41.9

❑ **Qualitative comparison with BaS-Net (AAAI'20)**