Background Suppression Network for Weakly-supervised Temporal Action Localization

Pilhyeon Lee[1] Youngjung Uh[2] Hyeran Byun[1*]

[1]Yonsei University, Seoul, Korea [2]Clova AI Research, NAVER Corp.

GitHub

Weakly-supervised temporal action localization

❑ Goal

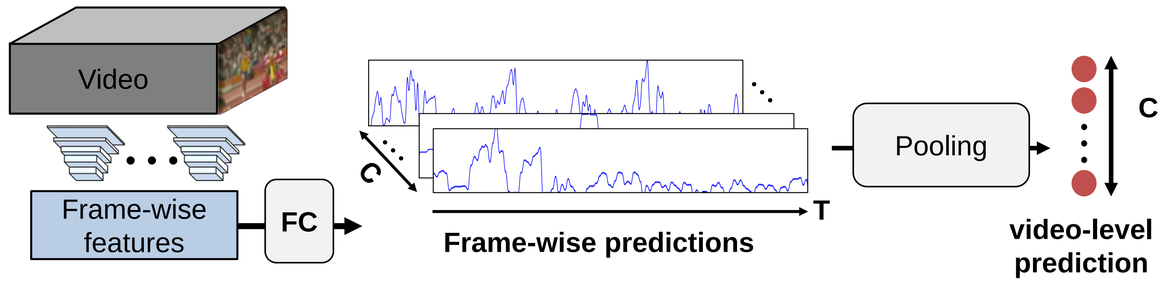
untrimmed videos → action intervals + classes
(analogy: images → bounding boxes + classes)

❑ Given

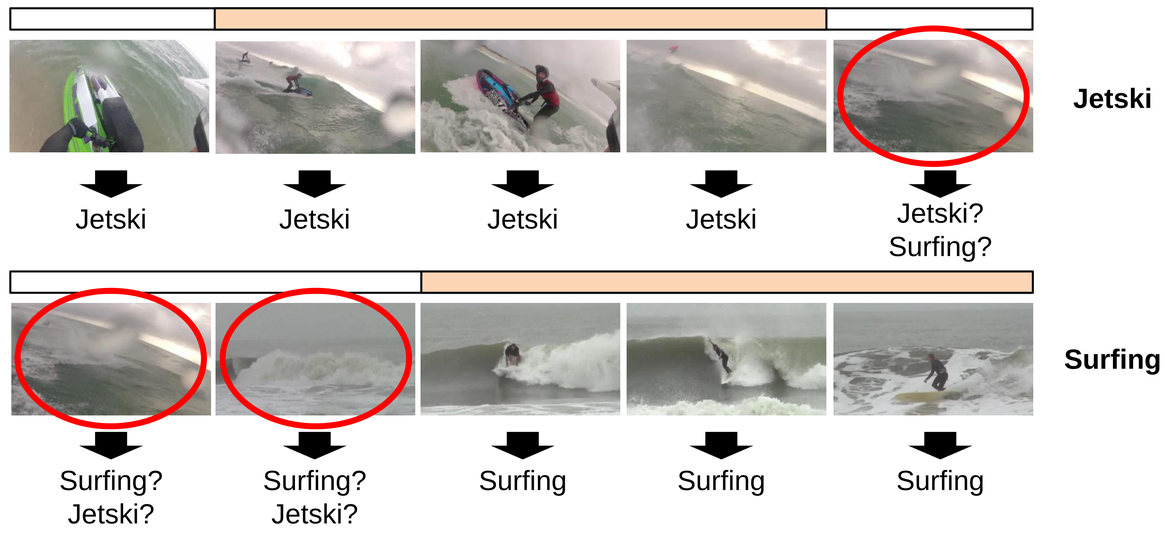
video-level labels
(not frame-level, analogy: image-level labels)

Common practice and pitfall

❑ Framework



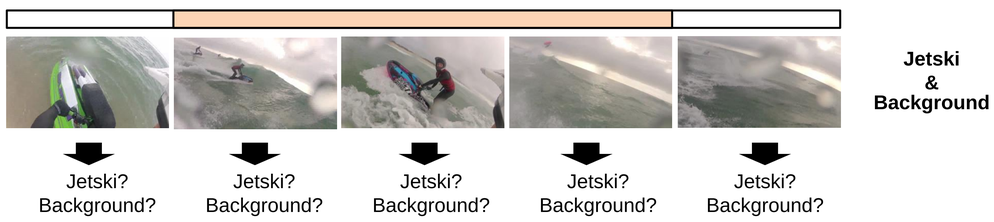
❑ Pitfall



Proposed Method

❑ Motivation

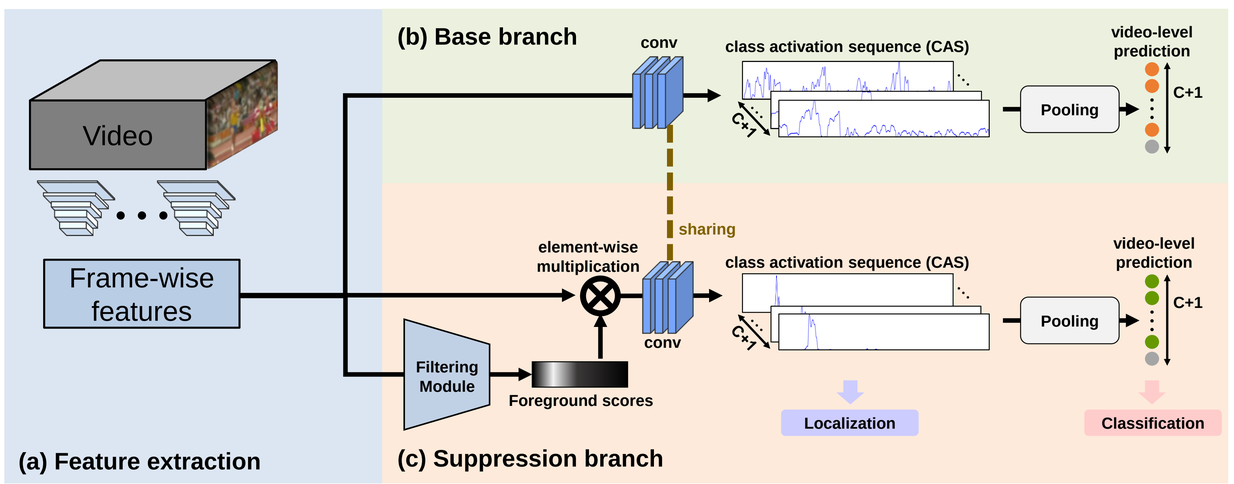
- ➢ Background frames should be classified as background class
- ➢ All untrimmed videos contain background frames



❑ Two-branch weight-sharing strategy

- ➢ Base & Suppression branch share weights
- ➢ Base branch learns
  - ✓ All videos = original class + background class
- ➢ Suppression branch learns
  - ✓ All videos = original class + NOT background class
  - ✓ Filtering module suppresses background frames

❑ BaS-Net architecture



Experimental Results

❑ Ablation study on THUMOS'14

	Base branch	Background class	Suppression branch	mAP@IoU 0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	average
Baseline	√			32.3	25.2	19.8	15.9	12.0	8.7	4.7	1.4	0.2	13.4
Base branch	√	√		28.5	23.0	18.1	13.7	9.2	5.8	2.7	0.8	0.1	11.3
Suppression branch		√	√	49.1	42.5	33.5	26.0	18.6	12.8	6.2	2.0	**0.5**	21.2
BaS-Net	√	√	√	**58.2**	**52.3**	**44.6**	**36.0**	**27.0**	**18.6**	**10.4**	**3.9**	**0.5**	**27.9**

Action Localization

	F-measure
Base branch	0.541
Suppression branch	0.775
BaS-Net	0.846

Background Detection

❑ 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	0.8	0.9
Full	TAL-Net (2018)	59.8	57.1	53.2	**48.5**	**42.8**	**33.8**	**20.8**	-	-
	Action Search (2018)	51.8	42.4	30.8	20.2	11.1	-	-	-	-
	BSN (2018)	-	-	53.5	45.0	36.9	28.4	20.0	-	-
	GTAN (2019)	**69.1**	**63.7**	**57.8**	47.2	38.8	-	-	-	-
Weak†	STAR (2019)	**68.8**	**60.0**	**48.7**	**34.7**	**23.0**	-	-	-	-
Weak	UntrimmedNet (2017)	44.4	37.7	28.2	21.1	13.7	-	-	-	-
	Hide-and-seek (2017)	36.4	27.8	19.5	12.7	6.8	-	-	-	-
	STPN (UNT) (2018)	45.3	38.8	31.1	23.5	16.2	9.8	5.1	2.0	0.3
	AutoLoc (2018)	-	-	35.8	29.0	21.2	13.4	5.8	-	-
	W-TALC (UNT) (2018)	49.0	42.8	32.0	26.0	18.8	-	6.2	-	-
	Liu et al. (UNT) (2019)	53.5	46.8	37.5	29.1	19.9	12.3	6.0	-	-
	Ours (UNT)	**56.2**	**50.3**	**42.8**	**34.7**	**25.1**	**17.1**	**9.3**	**3.7**	**0.5**
	STPN (I3D) (2018)	52.0	44.7	35.5	25.8	16.9	9.9	4.3	1.2	0.1
	W-TALC (I3D) (2018)	55.2	49.6	40.1	31.1	22.8	-	7.6	-	-
	MAAN (2019)	**59.8**	50.8	41.1	30.6	20.3	12.0	6.9	2.6	0.2
	Liu et al. (I3D) (2019)	57.4	50.8	41.2	32.1	23.1	15.0	7.0	-	-
	Ours (I3D)	58.2	**52.3**	**44.6**	**36.0**	**27.0**	**18.6**	**10.4**	**3.9**	**0.5**

THUMOS'14

Supervision	Method	mAP@IoU 0.5	0.75	0.95	average
Full	TAL-Net (2018)	38.2	18.3	1.3	20.2
	BSN (2018)	52.5	33.5	**8.9**	33.7
	GTAN (2019)	**52.6**	**34.1**	**8.9**	**34.3**
Weak†	STAR (2019)	**31.1**	**18.8**	**4.7**	-
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
	Ours	**34.5**	**22.5**	4.9	**22.2**

ActivityNet 1.3

Supervision	Method	mAP@IoU 0.5	0.75	0.95	average
Full	SSN (2017)	**41.3**	**27.0**	**6.1**	**26.6**
Weak	AutoLoc (2018)	27.3	15.1	3.3	16.0
	W-TALC (2018)	37.0	-	-	18.0
	Liu et al. (2019)	36.8	22.0	**5.6**	22.4
	Ours	**38.5**	**24.2**	**5.6**	**24.3**

ActivityNet 1.2

❑ Qualitative results

(a) An example of *VolleyballSpiking* action

(b) An example of *Shotput* action

(c) An example of *SoccerPenalty* action