Learning Action Completeness from Points for Weakly-supervised Temporal Action Localization

Pilhyeon Lee[1] Hyeran Byun[1,2*]

[1]Department of Computer Science, Yonsei University

[2]Graduate School of Artificial Intelligence, Yonsei University

GitHub

2021 ICCV OCTOBER 11-17 VIRTUAL

Check out our paper for more information.
https://arxiv.org/abs/2108.05029

Problem

- **Training phase**

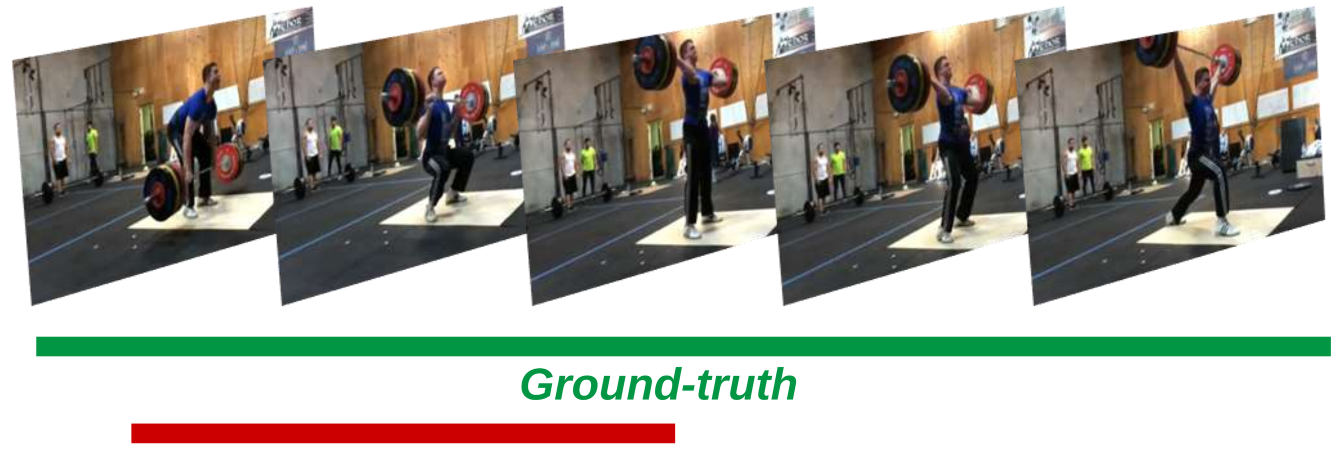
The cost-effective **point-level** labels are utilized for training.
(45s for video-level *vs.* 50s for point-level *vs.* 300s for full sup. per 1-min video)

- **Test phase**

At inference time, the model should predict the **temporal intervals** as well as the **classes** of action instances.

Motivation

Despite the excellent performance in spotting actions, existing works **fail to learn action completeness** due to the discontinuous nature of points, leading to fragmentary predictions. (e.g., IoUs $\leq$ 0.4)



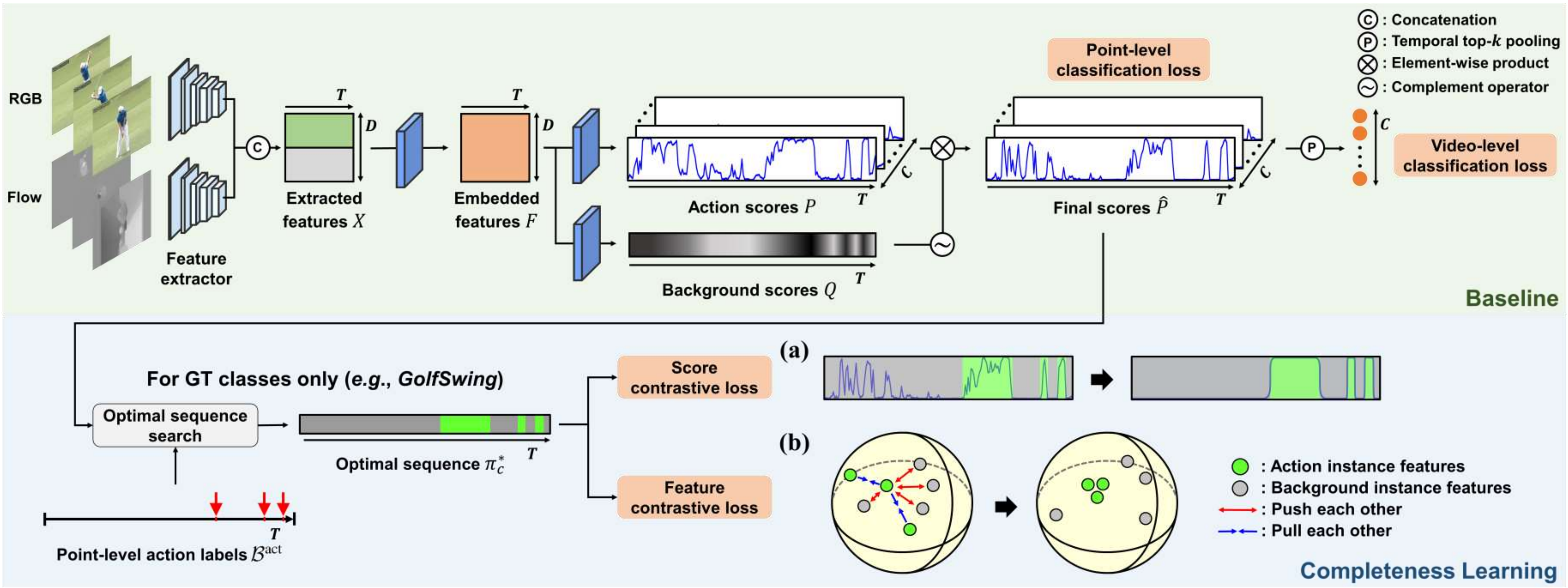
For instance, a model might detect only a sub-action, e.g., "*Power Clean*", rather than the full extent of "*Clean and Jerk*".

Goal

To tackle the challenge, we propose to generate dense pseudo labels and explicitly provide **guidance** to the model for action completeness learning.

Method

Our model consists of (top) the baseline part and (bottom) the completeness learning part.



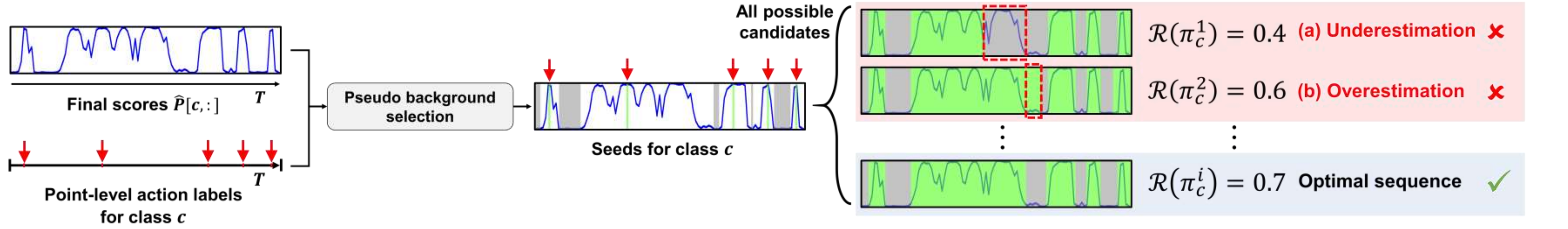
- **Baseline**

It learns from the common video- and point-level classification losses. At this time, we additionally select pseudo background points to supplement action ones.

$$\mathcal{L}_{\text{baseline}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{point}}$$

- **Completeness Learning**

To allow the model to learn action completeness, we first search for the optimal sequence that is likely to contain complete action instances, while avoiding under- and over-estimation cases.



To learn action completeness from the obtained optimal sequence, we design two loss functions that contrast action instances from background ones in terms of action scores and feature similarities.

(1) The **score contrastive loss** encourages the model output to fit the optimal sequence.

$$\mathcal{L}_{\text{score}} = \frac{1}{\sum_{c=1}^{C} y^{\text{vid}}[c]} \sum_{c=1}^{C} y^{\text{vid}}[c] \left(1 - \mathcal{R}(\pi_c^*)\right)^{\beta}$$

(2) The **feature contrastive loss** encourages action features to attract each other but to repel background ones.

$$\mathcal{L}_{\text{feat}} = \frac{1}{\sum_{c=1}^{C} \mathbb{1}\left[\sum_{n=1}^{N_c} z_n^c > 1\right]} \sum_{c=1}^{C} \mathbb{1}\left[\sum_{n=1}^{N_c} z_n^c > 1\right] \frac{-1}{\sum_{n=1}^{N_c} z_n^c} \sum_{n=1}^{N_c} z_n^c \log \frac{\sum_{\forall o \neq n} z_o^c \exp(\tilde{f}_n^c \cdot \tilde{f}_o^c / \tau)}{\sum_{\forall m \neq n} \exp(\tilde{f}_n^c \cdot \tilde{f}_m^c / \tau)}$$

Experiments

Our model largely surpasses the weakly-supervised state-of-the-arts, even performing favorably against fully-supervised counterparts at the **6× cheaper cost**.

Supervision	Method	mAP@IoU (%) 0.1	0.2	0.3	0.4	0.5	0.6	0.7	AVG (0.1:0.5)	AVG (0.3:0.7)
Frame-level (Full)	BMN [26]	-	-	56.0	47.4	38.8	29.7	20.5	-	38.5
	P-GCN [67]	69.5	67.8	63.6	57.8	49.1	-	-	61.6	-
	G-TAD [61]	-	-	54.5	47.6	40.2	30.8	23.4	-	39.3
	BC-GNN [1]	-	-	57.1	49.1	40.4	31.2	23.1	-	40.2
	Zhao *et al.* [71]	-	-	53.9	50.7	45.4	38.0	28.5	-	43.3
Video-level (Weak)	Lee *et al.* [22]	67.5	61.2	52.3	43.4	33.7	22.9	12.1	51.6	32.9
	CoLA [69]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1
	AUMN [33]	66.2	61.9	54.9	44.4	33.3	20.5	9.0	52.1	32.4
	TS-PCA [30]	67.6	61.1	53.4	43.4	34.3	24.7	13.7	52.0	33.9
	UGCT [64]	69.2	62.9	55.5	46.5	35.9	23.8	11.4	54.0	34.6
Point-level (Weak)	SF-Net† [35]	71.0	63.4	53.2	40.7	29.3	18.4	9.6	51.5	30.2
	Ju *et al.*† [14]	72.8	64.9	58.1	46.4	34.5	21.8	11.9	55.3	34.5
	Ours†	**75.1**	**70.5**	**63.3**	**55.2**	**43.9**	**33.3**	**20.8**	**61.6**	**43.3**
	Moltisanti *et al.*‡ [42]	24.3	19.9	15.9	12.5	9.0	-	-	16.3	-
	SF-Net‡ [35]	68.3	62.3	52.8	42.2	30.5	20.6	12.0	51.2	31.6
	Ju *et al.*‡ [14]	72.3	64.7	58.2	47.1	35.9	23.0	12.8	55.6	35.4
	Ours‡	**75.7**	**71.4**	**64.6**	**56.5**	**45.3**	**34.5**	**21.8**	**62.7**	**44.5**

We validate that our action completeness learning indeed helps in detecting complete action instances (See the improvements in **mAP@0.5** and **0.7**).

$\mathcal{L}_{\text{video}}$	$\mathcal{L}_{\text{point}}$	$\mathcal{L}_{\text{score}}$	$\mathcal{L}_{\text{feat}}$	mAP@IoU (%) 0.1	0.3	0.5	0.7	AVG
✓	✗	✗	✗	51.9	37.1	20.3	6.0	28.7
✓	✓	✗	✗	70.7	58.1	40.7	16.1	47.3
✓	✓	✓	✗	75.1	64.4	44.5	20.0	52.0
✓	✓	✗	✓	72.1	60.5	42.1	17.9	49.0
✓	✓	✓	✓	**75.7**	**64.6**	**45.3**	**21.8**	**52.8**

It is clearly shown by the visualization results that our model produces more **complete** action predictions (IoUs > 0.6).