2021 ICCV OCTOBER 11-17 VIRTUAL

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

Oral presentation, ICCV 2021

Pilhyeon Lee
Ph.D. student

Hyeran Byun
Professor

YONSEI UNIVERSITY

scan me

Temporal Action Localization

Goal: to predict the temporal intervals of action instances.

Temporal Action Localization

BaseballPitch
(0' 49" ~ 0' 53")

BaseballPitch
(1' 59" ~ 2' 03")

Despite its great importance in video understanding, the heavy annotation cost limits its scalability.
(e.g., it takes 300 *sec* to annotate a 1-min video)

To bypass the high labeling cost, we focus on weak supervision.

Weakly-supervised Temporal Action Localization

Video-level: ***BaseballPitch***

The cheapest one is in the video-level, which indicates the presence (absence) of action classes. It takes 45 *sec* per 1-min video.

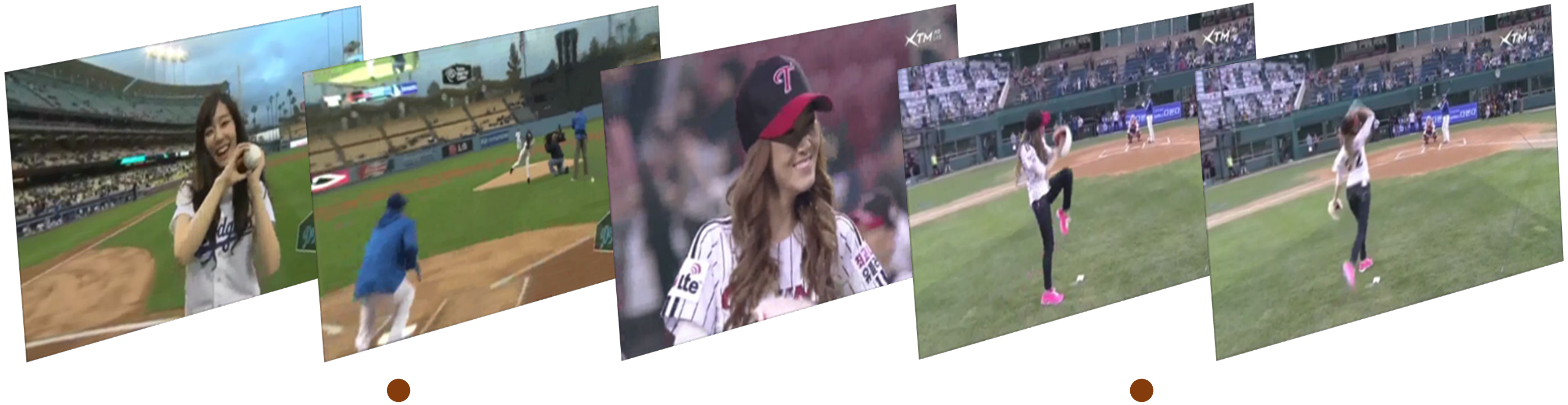
Weakly-supervised Temporal Action Localization

Video-level: *BaseballPitch*

Unfortunately, there is no free lunch.
The *cheaper* the annotation is, the *poorer* the model performs.

E.g., Bottom-Up$_{\text{[ECCV'20]}}$ 45.4% *vs.* EM-MIL$_{\text{[ECCV'20]}}$ 30.5%
(mAP@IoU=0.5)

Weakly-supervised Temporal Action Localization



BaseballPitch
(0' 52")

BaseballPitch
(2' 0")

Point-level (or single-frame) supervision has been proposed to bridge the gap.

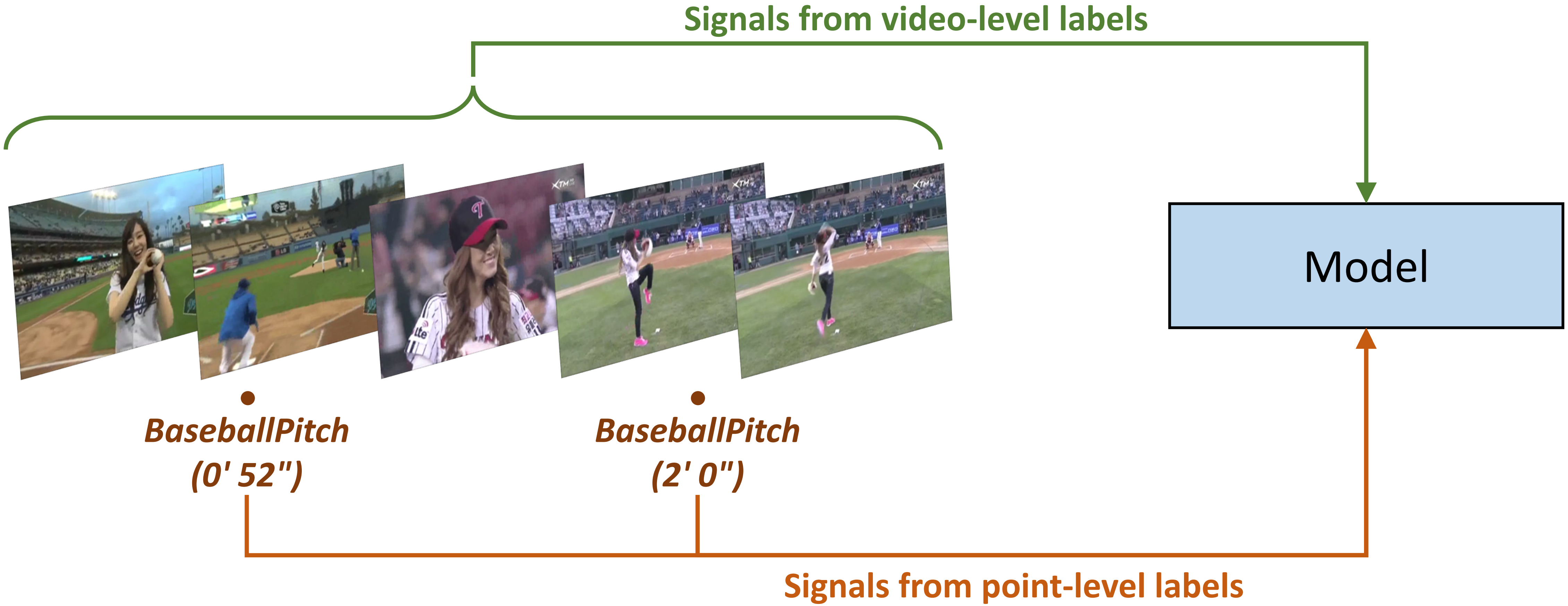
Weakly-supervised Temporal Action Localization

BaseballPitch
(0' 52")

BaseballPitch
(2' 0")

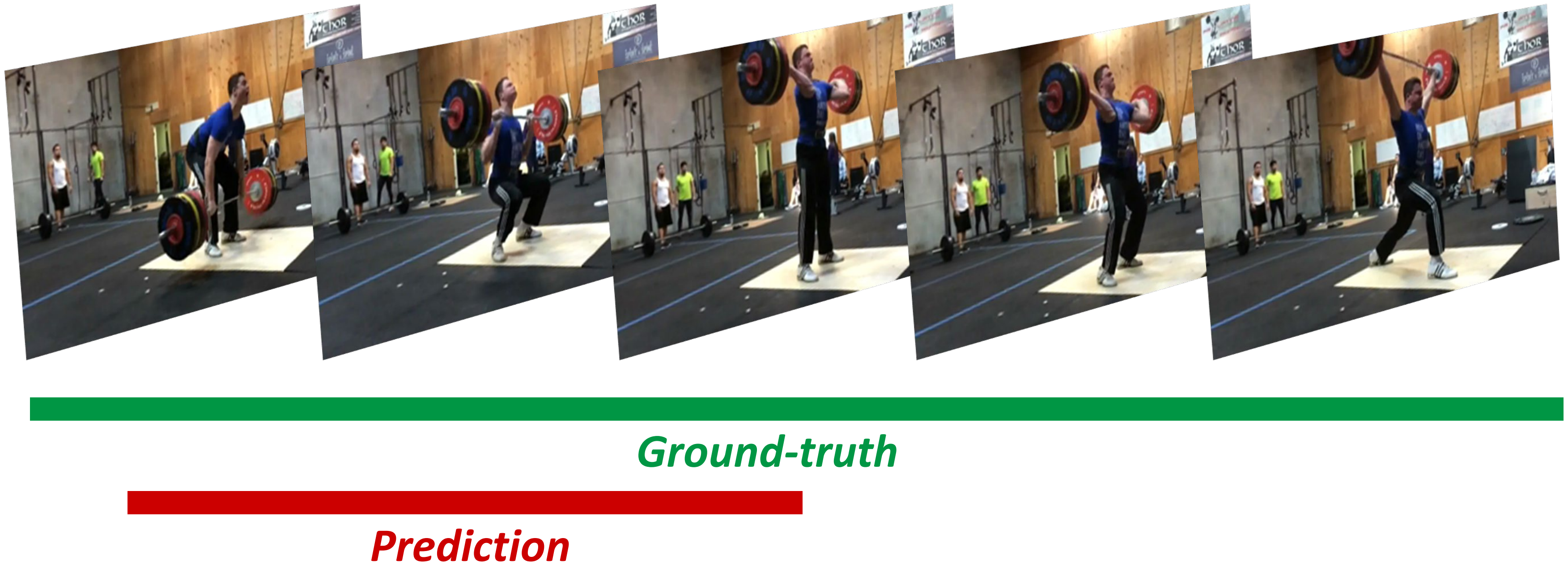
It avoids the rewind stage, and therefore has a comparable cost, e.g., 45 *sec* vs. 50 *sec.*
Meanwhile, it offers **far richer information**, e.g., action count and rough action locations.

Challenges of Prior Arts



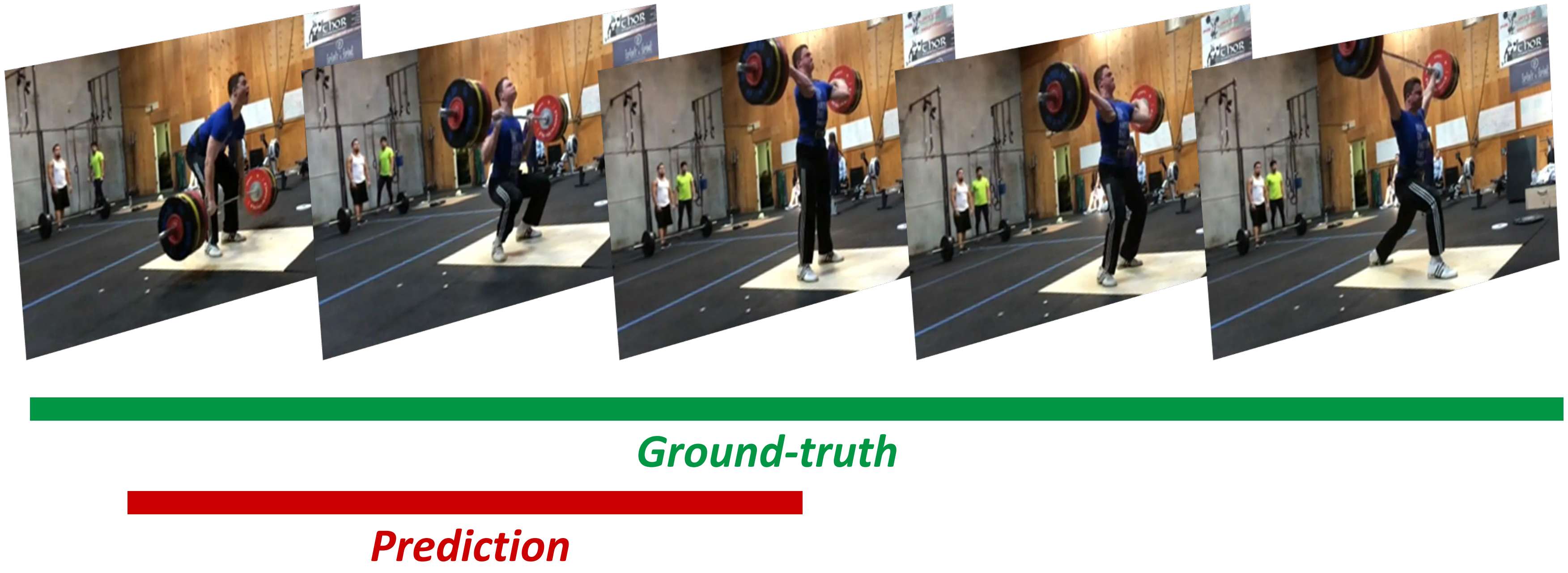
Previous methods simply learn from video- and point-level supervision.

Challenges of Prior Arts



While point-level supervision helps the models to spot action instances (low IoUs), they fail to learn *action completeness* due to the discontiguous property of points.

Challenges of Prior Arts



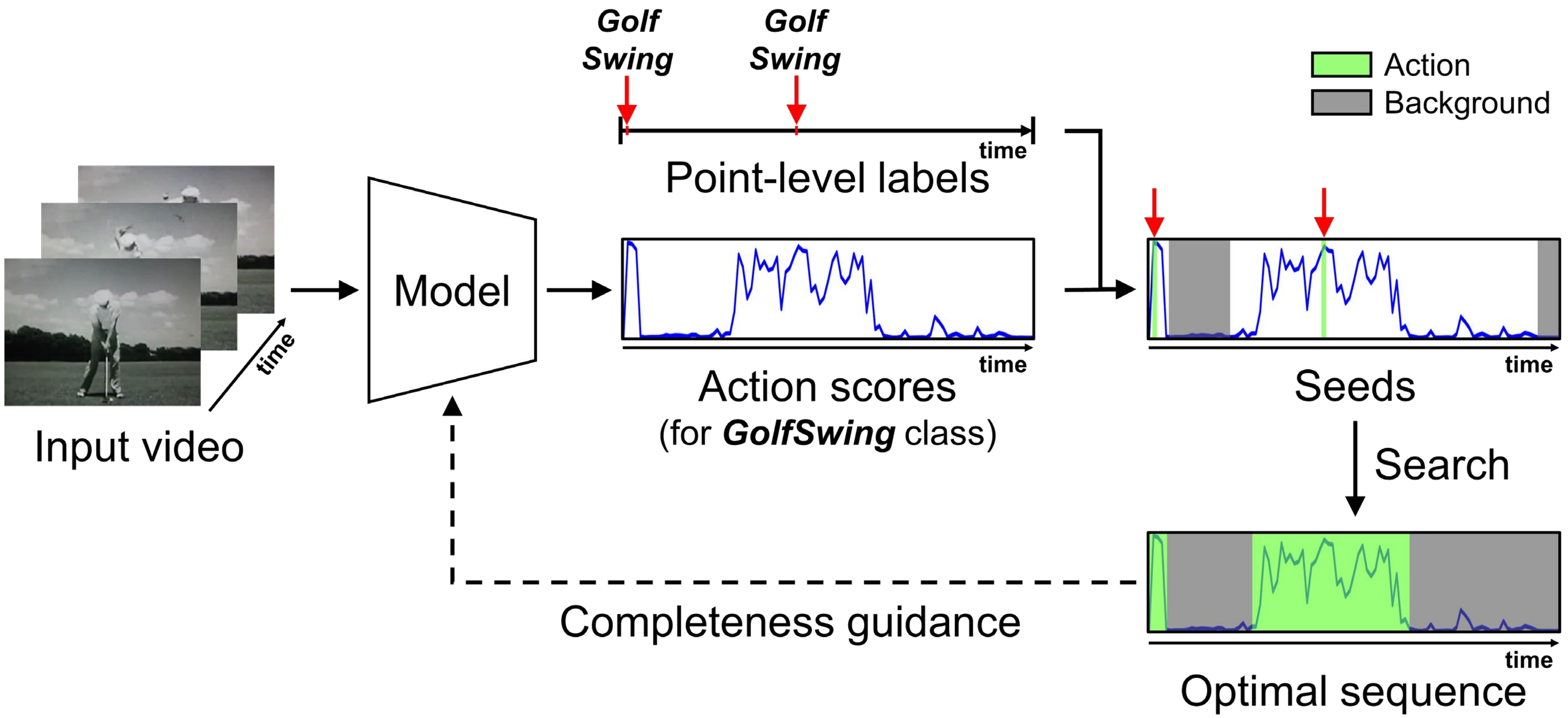
While point-level supervision helps the models to spot action instances (low IoUs), they fail to learn *action completeness* due to the discontiguous property of points.

➔ We propose to explicitly learn *action completeness* from points.

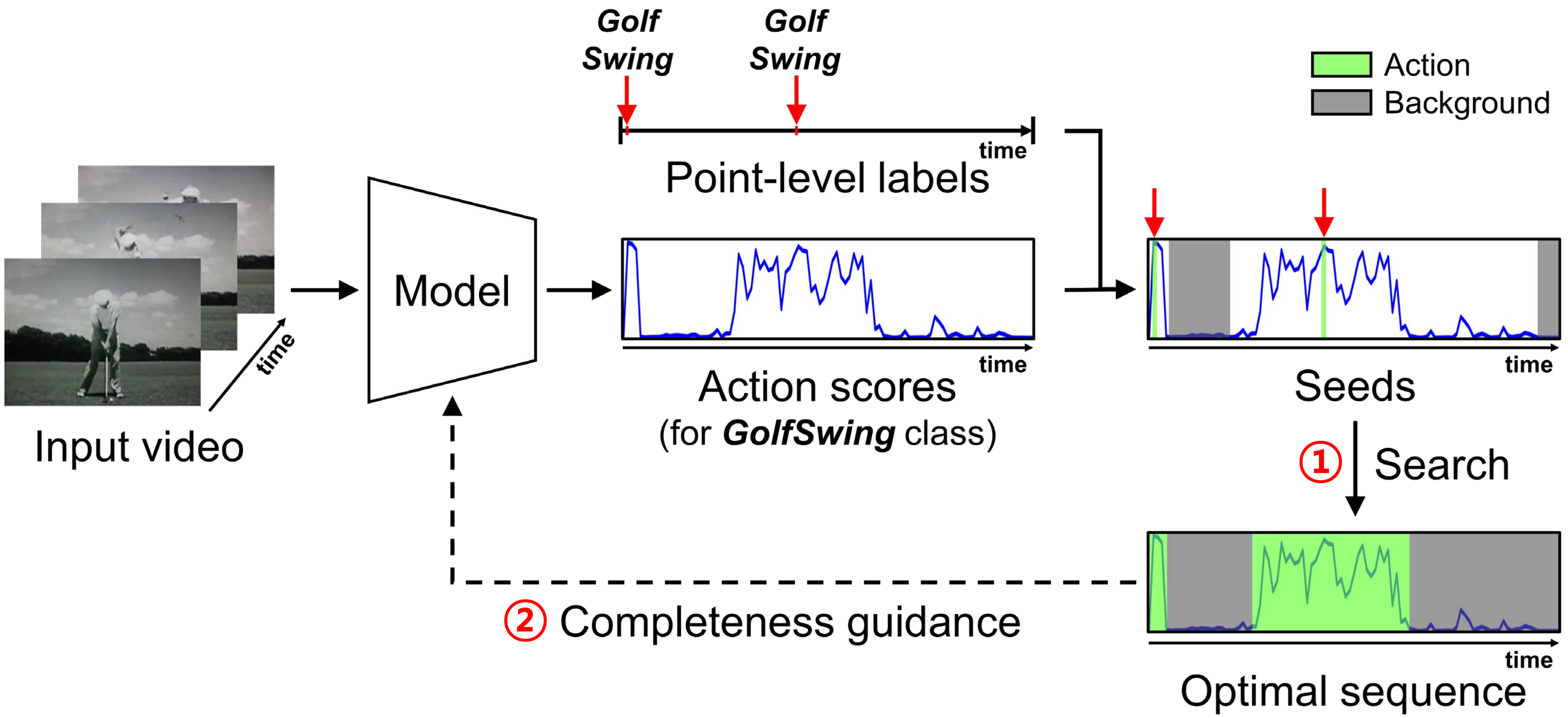
Our idea is simple.

If continuity is the key, why don't we generate dense pseudo labels that can provide *completeness guidance* for the model?

Method


Golf Swing
Golf Swing
time
Point-level labels
Action
Background
Model
Input video
time
Action scores
(for GolfSwing class)
time
Seeds
time
Search
Optimal sequence
time
Completeness guidance

Method



There remain two questions.

① How can we obtain the sequence that best suits the (unknown) ground truth?

② How can we effectively lead the model to learn action completeness?

How can we obtain the sequence that best suits the (unknown) ground truth?

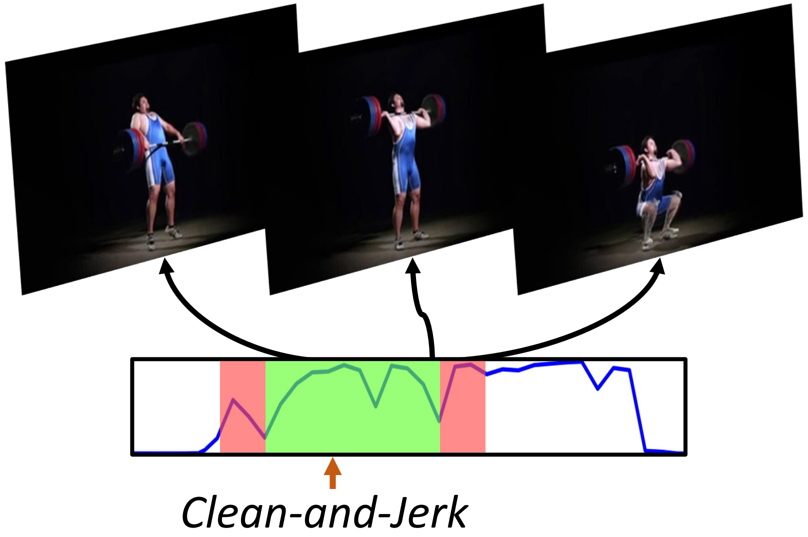
Method

① How can we obtain the sequence that best suits the (unknown) ground truth?

We utilize the score contrast between inner and outer regions (–)
as a *proxy* to judge the degree of action completeness for a predicted instance.

Method

① How can we obtain the sequence that best suits the (unknown) ground truth?

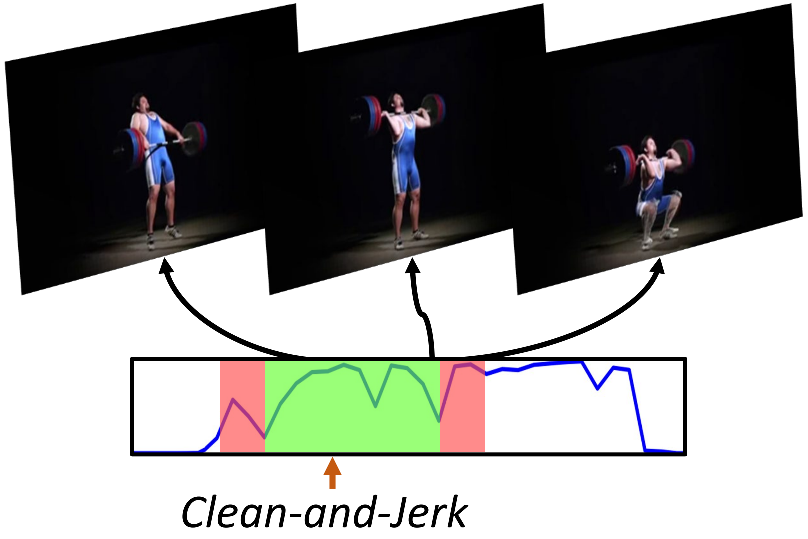


Underestimation
(*low* contrast)

We utilize the score contrast between inner and outer regions (–)
as a *proxy* to judge the degree of action completeness for a predicted instance.

Method

① How can we obtain the sequence that best suits the (unknown) ground truth?



Underestimation
(*low* contrast)

Overestimation
(*low* contrast)

We utilize the score contrast between inner and outer regions (–)
as a *proxy* to judge the degree of action completeness for a predicted instance.

Method

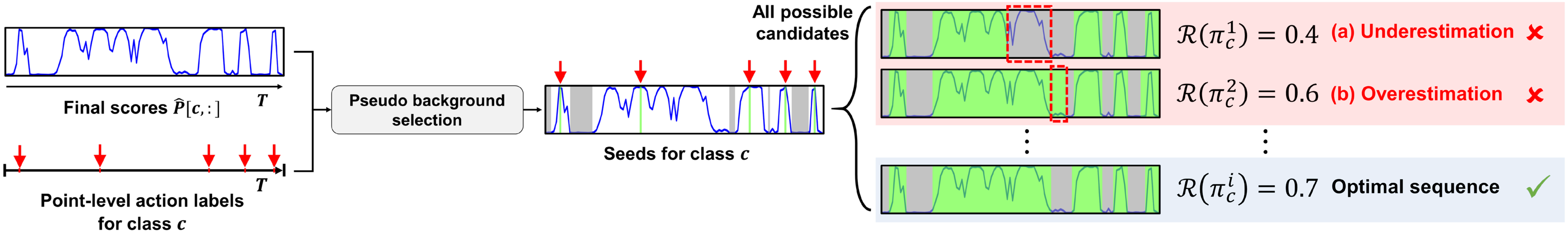
① How can we obtain the sequence that best suits the (unknown) ground truth?

Underestimation (*low* contrast) | Overestimation (*low* contrast) | Best fit (*high* contrast)

We utilize the score contrast between inner and outer regions (–)
as a *proxy* to judge the degree of action completeness for a predicted instance.

Method

① How can we obtain the sequence that best suits the (unknown) ground truth?



$$\mathcal{R}(\pi_c) = \frac{1}{N_c}\sum_{n=1}^{N_c}\Bigg(\underbrace{\frac{1}{l_n^c}\sum_{t=s_n^c}^{e_n^c} u_n^c(t)}_{\text{Inner score}} - \underbrace{\frac{1}{\lceil \delta l_n^c \rceil + \lfloor \delta l_n^c \rfloor}\Bigg(\sum_{t=s_n^c-\lceil \delta l_n^c \rceil}^{s_n^c-1} u_n^c(t) + \sum_{t=e_n^c+1}^{e_n^c+\lfloor \delta l_n^c \rfloor} u_n^c(t)\Bigg)}_{\text{Outer score}}\Bigg),$$

$$\text{where } u_n^c(t) = \begin{cases} \hat{p}_t[c], & \text{if } z_n^c = 1. \\ 1 - \hat{p}_t[c], & \text{otherwise.} \end{cases},$$

Our goal is to search for the optimal sequence: $\pi_c^* = \arg\max_{\pi_c} \mathcal{R}(\pi_c)$

How can we effectively lead the model to learn action completeness?

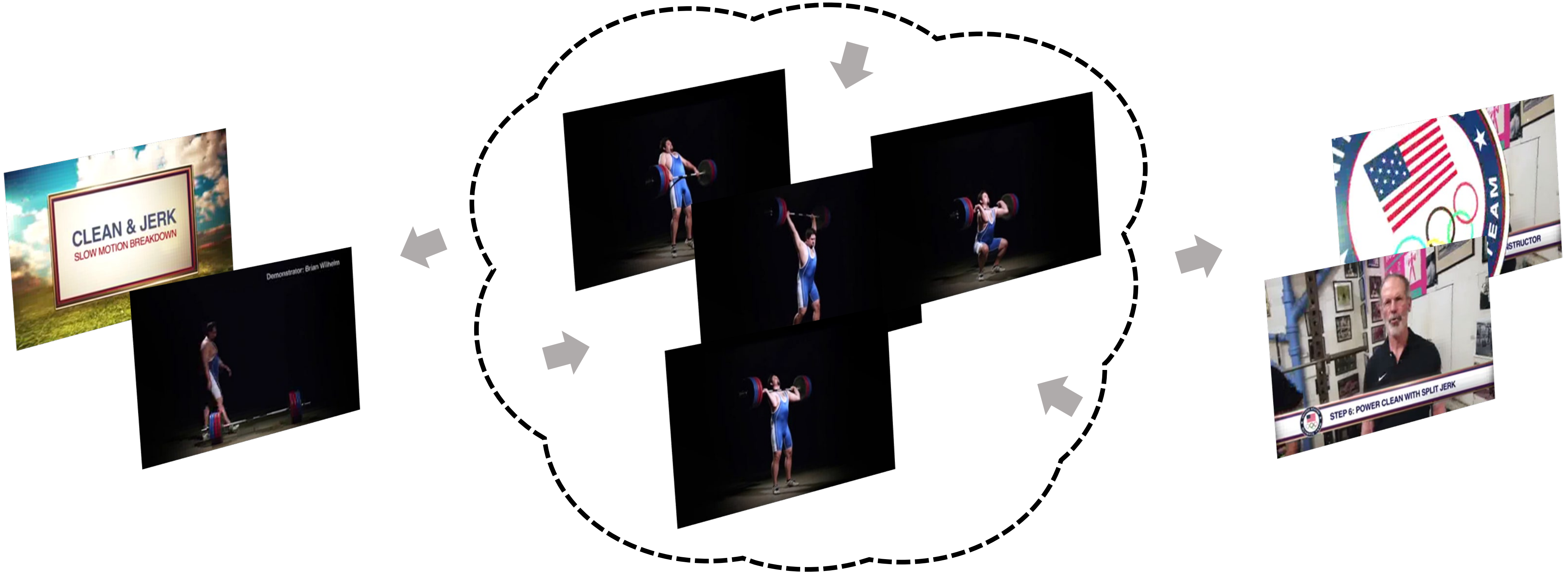
Method

② How can we effectively lead the model to learn action completeness?

We encourage the model to contrast action instances from their surrounding backgrounds.

Method

② How can we effectively lead the model to learn action completeness?



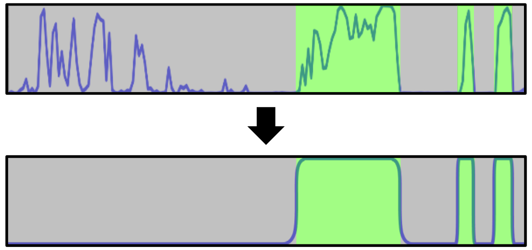
This instance-level contrastive strategy brings two advantages simultaneously, *i.e., intra-action compactness* and *action-background separation*.

Method

② How can we effectively lead the model to learn action completeness?

1) Score contrastive loss

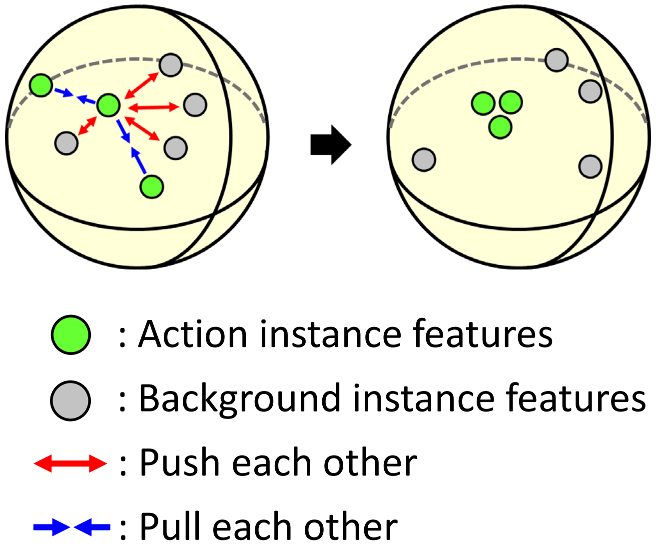
$$\mathcal{L}_{\text{score}} = \frac{1}{\sum_{c=1}^{C} y^{\text{vid}}[c]} \sum_{c=1}^{C} y^{\text{vid}}[c]\big(1 - \mathcal{R}(\pi_c^*)\big)^{\beta}$$



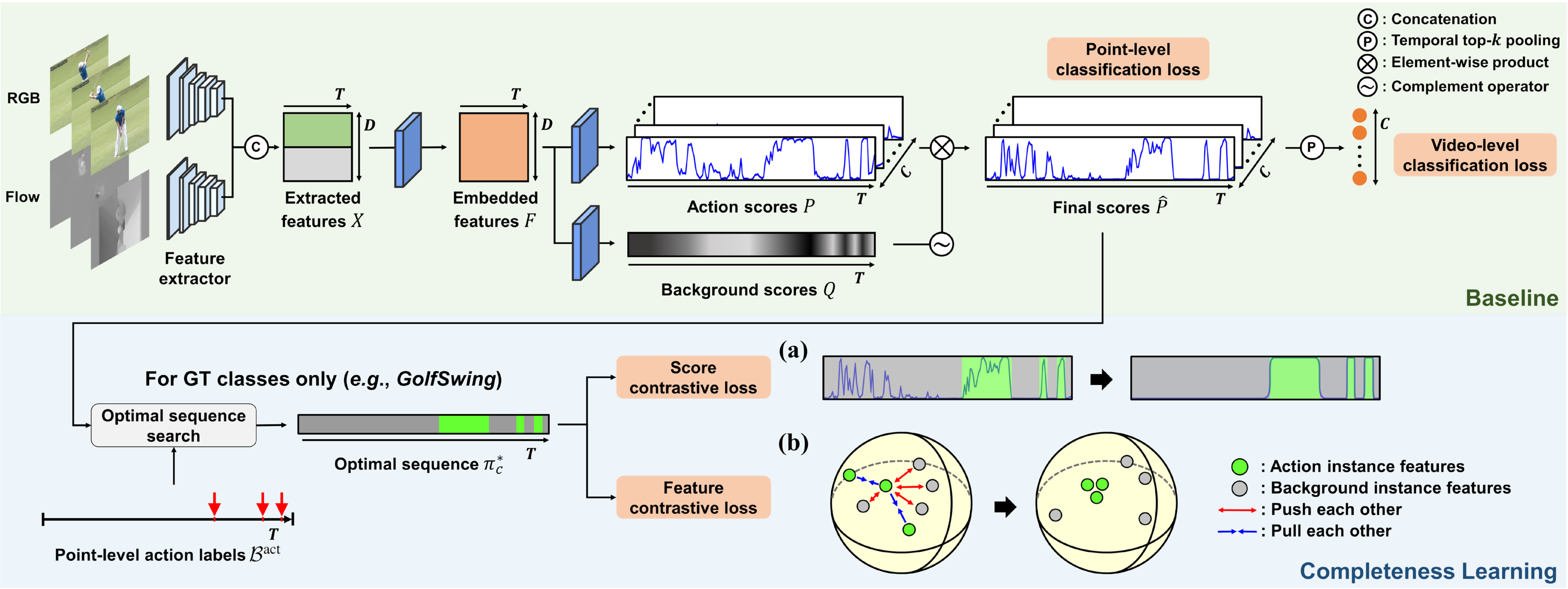
2) Feature contrastive loss

$$\mathcal{L}_{\text{feat}} = \frac{1}{\sum_{c=1}^{C} \mathbb{1}\left[\sum_{n=1}^{N_c} z_n^c > 1\right]} \sum_{c=1}^{C} \mathbb{1}\left[\sum_{n=1}^{N_c} z_n^c > 1\right] \ell_{\text{feat}}^c,$$

$$\text{where} \quad \ell_{\text{feat}}^c = -\frac{1}{\sum_{n=1}^{N_c} z_n^c} \sum_{n=1}^{N_c} z_n^c \log \frac{\sum_{\forall o \neq n} z_o^c \exp(\bar{f}_n^c \cdot \bar{f}_o^c / \tau)}{\sum_{\forall m \neq n} \exp(\bar{f}_n^c \cdot \bar{f}_m^c / \tau)},$$

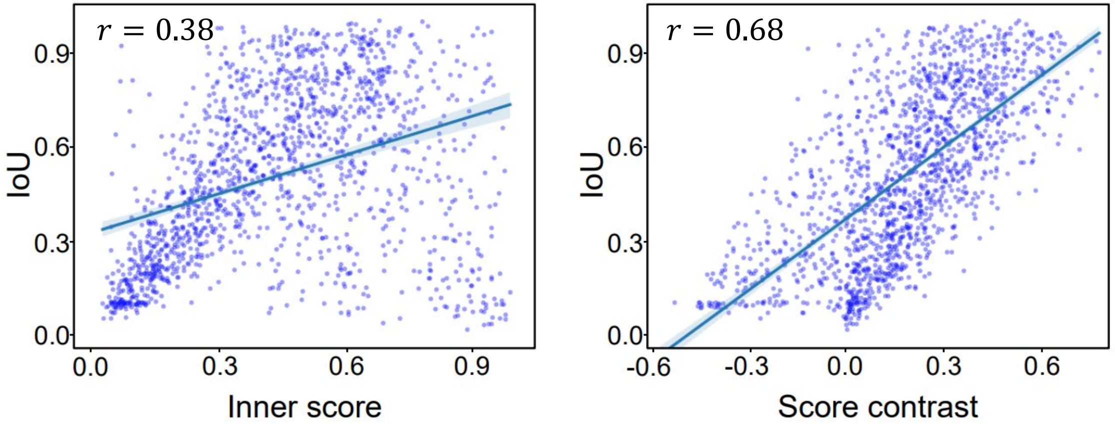


Method



RGB
Flow
Feature extractor
Extracted features X
Embedded features F
Action scores P
Background scores Q
Point-level classification loss
Final scores $\hat{P}$
Video-level classification loss
Ⓒ : Concatenation
Ⓟ : Temporal top-k pooling
⊗ : Element-wise product
~ : Complement operator
Baseline
For GT classes only (*e.g.*, *GolfSwing*)
Optimal sequence search
Optimal sequence π_c^*
Point-level action labels $\mathcal{B}^{\mathrm{act}}$
Score contrastive loss
Feature contrastive loss
(a)
(b)
: Action instance features
: Background instance features
: Push each other
: Pull each other
Completeness Learning

Analysis



inner score *vs.* score contrast

Scoring method	Sequence accuracy	mAP@IoU (%) 0.1	0.3	0.5	0.7	AVG
Baseline	N/A	70.7	58.1	40.7	16.1	47.3
(a) Inner scores	74.0	74.7	61.4	40.9	15.2	49.0
(b) Contrast-act	80.1	74.3	63.3	43.6	19.5	50.8
(c) Contrast-both	**83.9**	**75.7**	**64.6**	**45.3**	**21.8**	**52.8**

Comparison of scoring variants

How well does the score contrast represent the action completeness?

Analysis

$\mathcal{L}_{\text{video}}$	$\mathcal{L}_{\text{point}}$	$\mathcal{L}_{\text{score}}$	$\mathcal{L}_{\text{feat}}$	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**

Effect of each completeness guidance

Method	Distribution	Sequence accuracy	mAP@IoU (%) 0.3	0.5	0.7	AVG
SF-Net [35]	Manual	N/A	53.3	28.8	9.7	40.6
	Uniform	N/A	52.0	30.2	11.8	40.5
	Gaussian	N/A	47.4	26.2	9.1	36.7
Ju *et al.* [14]	Manual	N/A	58.1	34.5	11.9	44.3
	Uniform	N/A	55.6	32.3	12.3	42.9
	Gaussian	N/A	58.2	35.9	12.8	44.8
Ours	Manual	83.7	63.3	43.9	20.8	51.7
	Uniform	76.6	60.4	42.6	20.2	49.3
	Gaussian	**83.9**	**64.6**	**45.3**	**21.8**	**52.8**

Comparison of different label distributions

The action completeness learning indeed helps the model to localize more comprehensive action instances regardless of the point distributions.

State-of-the-art Comparison

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**

Results on THUMOS'14

State-of-the-art Comparison

Dataset	Method	mAP@IoU (%) 0.1	0.3	0.5	0.7	AVG
GTEA	SF-Net [35]	58.0	37.9	19.3	11.9	31.0
	SF-Net* [35]	52.9	37.6	21.7	13.7	31.1
	Ju *et al.* [14]	59.7	38.3	21.9	18.1	33.7
	Li *et al.* [24]	60.2	44.7	28.8	12.2	36.4
	Ours	**63.9**	**55.7**	**33.9**	**20.8**	**43.5**
BEOID	SF-Net [35]	62.9	40.6	16.7	3.5	30.9
	SF-Net* [35]	64.6	42.2	27.3	12.2	36.5
	Ju *et al.* [14]	63.2	46.8	20.9	5.8	34.9
	Li *et al.* [24]	71.5	40.3	20.3	5.5	34.4
	Ours	**76.9**	**61.4**	**42.7**	**25.1**	**51.8**

Results on GTEA & BEOID

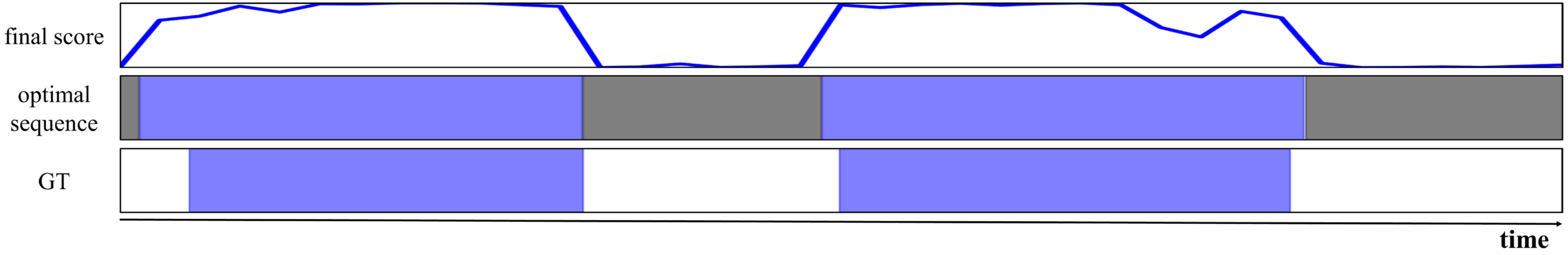
Supervision	Method	mAP@IoU (%) 0.5	0.75	0.95	AVG
Frame-level	SSN [72]	41.3	27.0	6.1	26.6
Video-level	Lee *et al.* [22]	41.2	25.6	6.0	25.9
	AUMN [33]	42.0	25.0	5.6	25.5
	UGCT [64]	41.8	25.3	5.9	25.8
	CoLA [69]	42.7	25.7	5.8	26.1
Point-level	SF-Net [35]	37.8	-	-	22.8
	Ours	**44.0**	**26.0**	**5.9**	**26.8**

Results on ActivityNet1.2

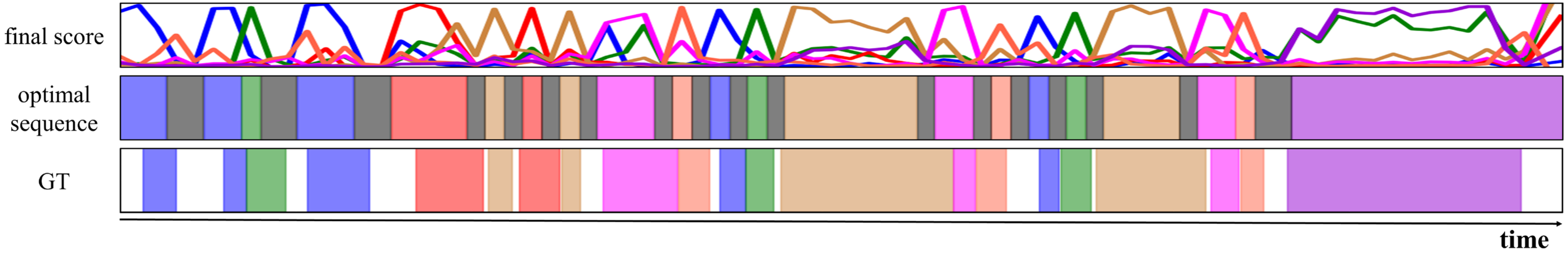
Supervision	Method	mAP@IoU (%) 0.5	0.75	0.95	AVG
Frame-level	BMN [26]	50.1	34.8	8.3	33.9
	P-GCN [67]	48.3	33.2	3.3	31.1
	G-TAD [61]	50.4	34.6	9.0	34.1
	BC-GNN [1]	50.6	34.8	9.4	34.2
	Zhao *et al.* [71]	43.5	33.9	9.2	30.1
Video-level	Lee *et al.* [22]	37.0	23.9	5.7	23.7
	AUMN [33]	38.3	23.5	5.2	23.5
	TS-PCA [64]	37.4	23.5	5.9	23.7
Point-level	Ours	**40.4**	**24.6**	**5.7**	**25.1**

Results on ActivityNet1.3

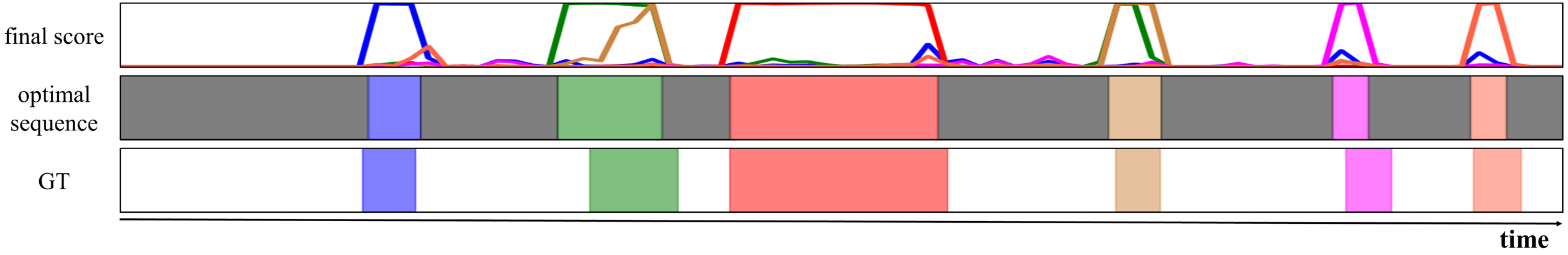
Optimal Sequence Visualization



(a) An example from THUMOS’14 (video_validation_0000261)

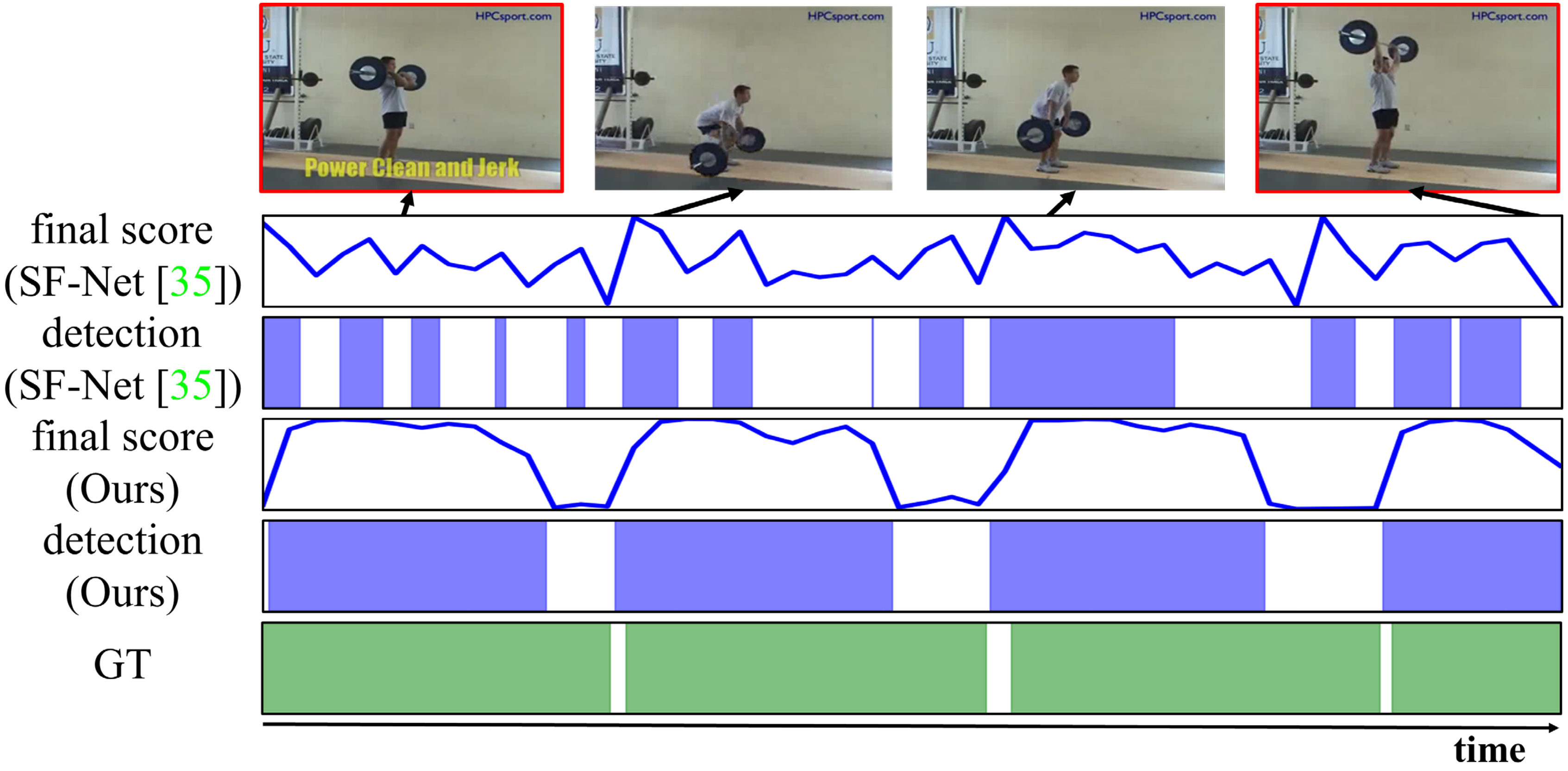


(b) An example from GTEA (S1_CofHoney_C1)



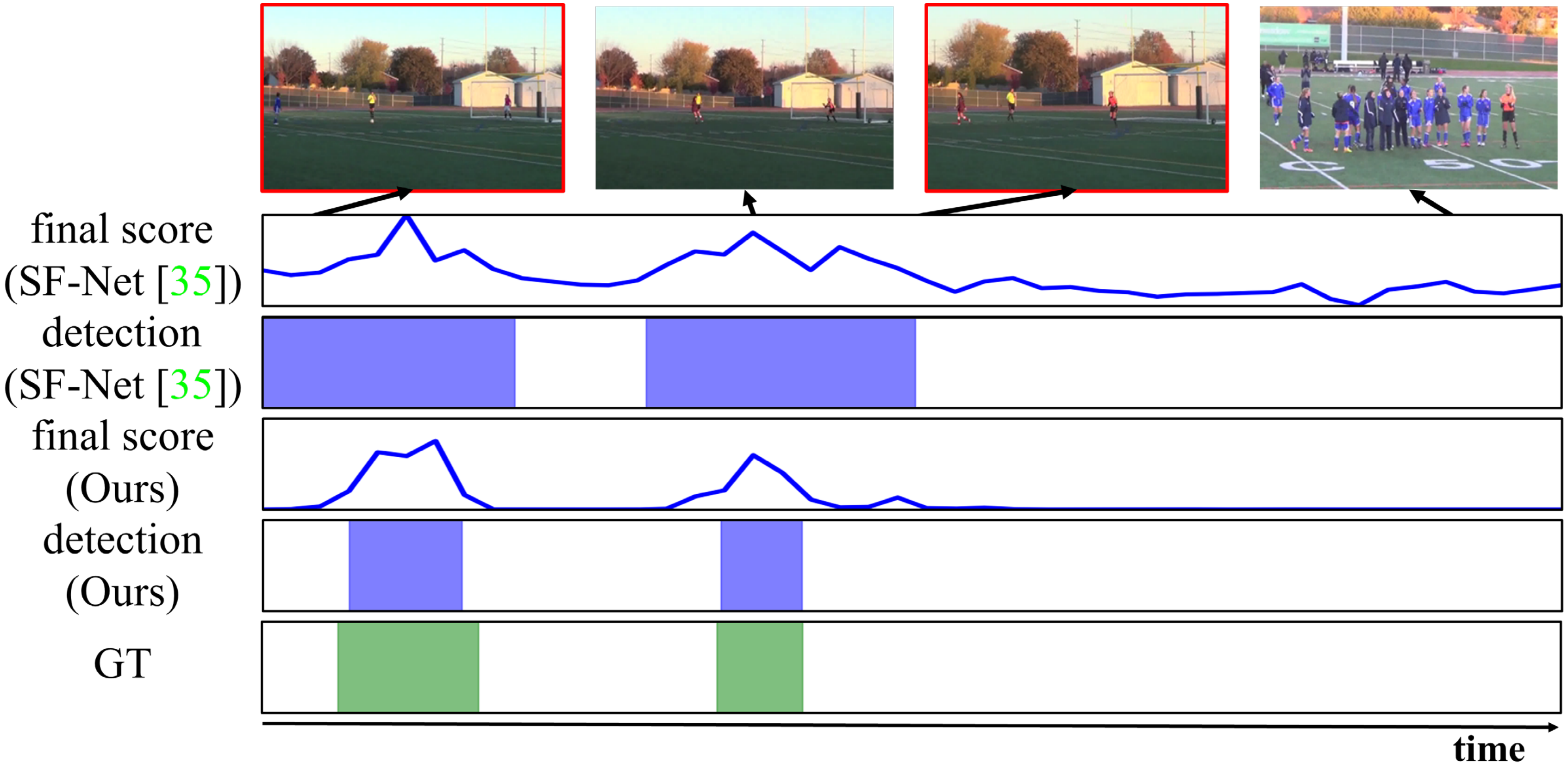
(c) An example from BEOID (02_Desk1)

Qualitative Comparison



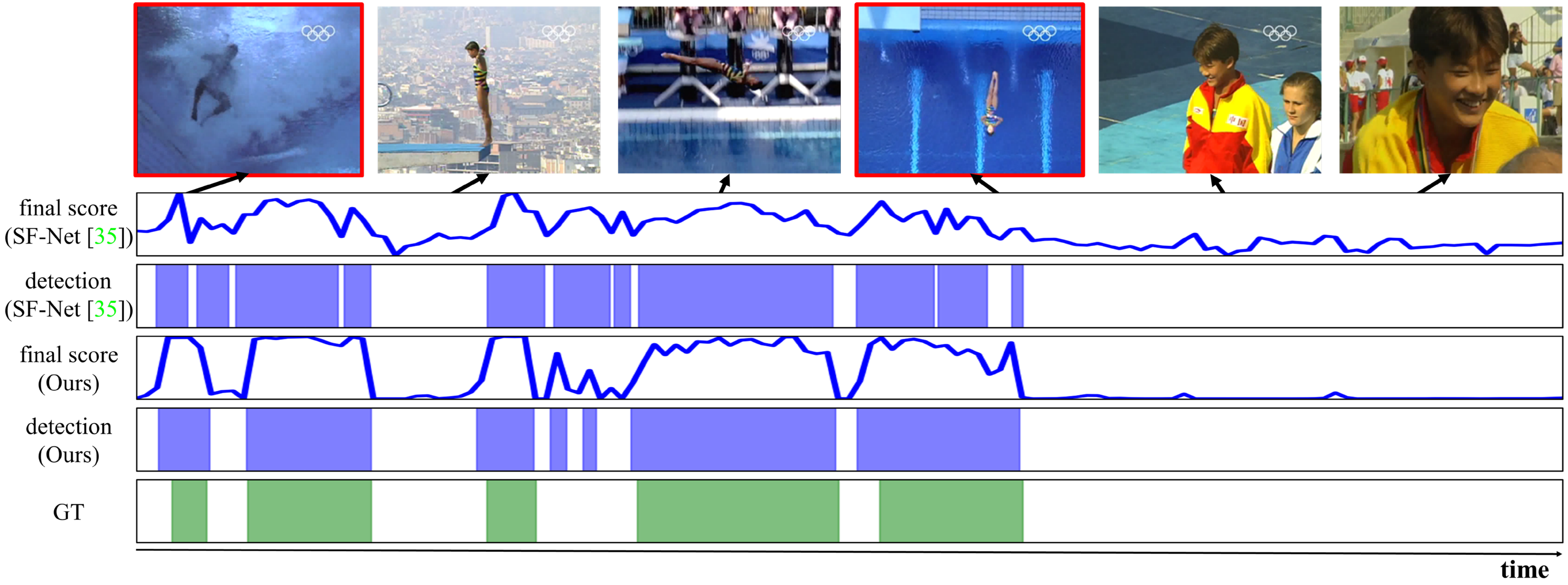
An example of *CleanAndJerk* action

Qualitative Comparison



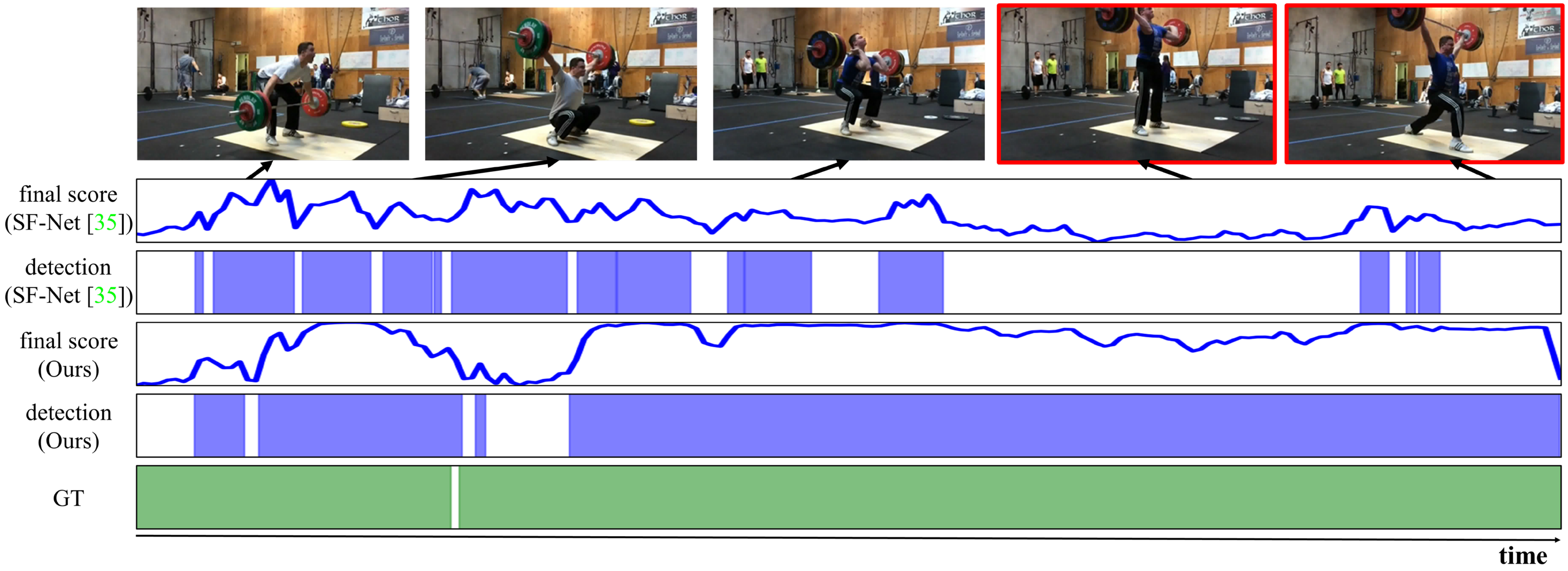
An example of *SocckerPenalty* action

Qualitative Comparison



(a) An example of *Diving* action (video_test_0001309)

Qualitative Comparison



(b) An example of *CleanAndJerk* action (video_test_000058)

Thank you!

Contact: lph1114@yonsei.ac.kr

scan me