25th International Conference on Pattern Recognition

ActionSpotter: Deep Reinforcement Learning Framework for Temporal Action Spotting in Videos

Guillaume VAUDAUX-RUTH[1,2], Adrien CHAN-HON-TONG[1,3], Catherine[2] ACHARD

[1]ONERA [2]Sorbonne Université [3]Université Paris-Saclay

ONERA THE FRENCH AEROSPACE LAB SORBONNE UNIVERSITÉ université PARIS-SACLAY

Problem setup

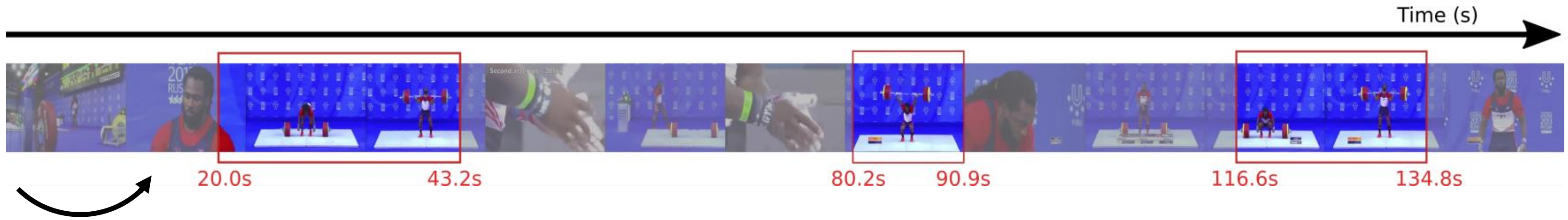
Action Spotting

Problem setup

Action Spotting

Problem setup

Action Spotting

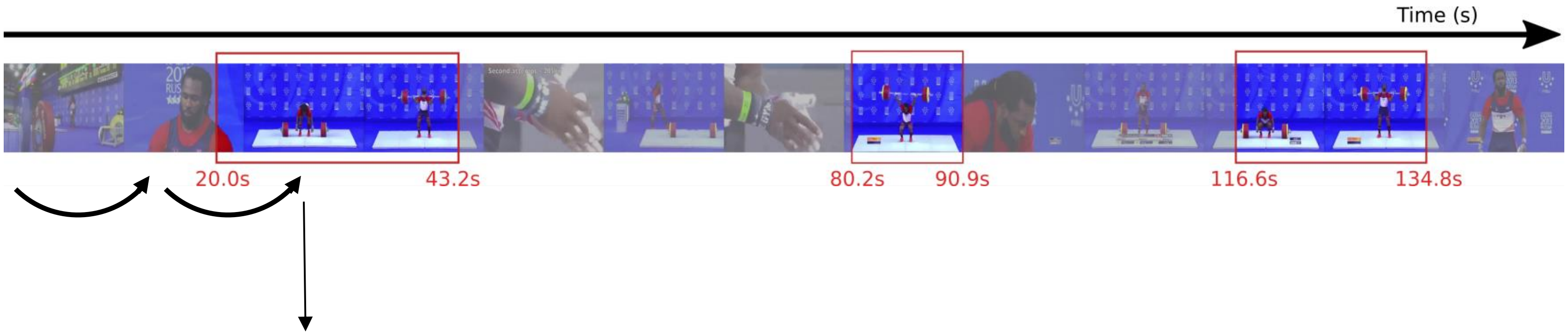


Problem setup

Action Spotting

Problem setup

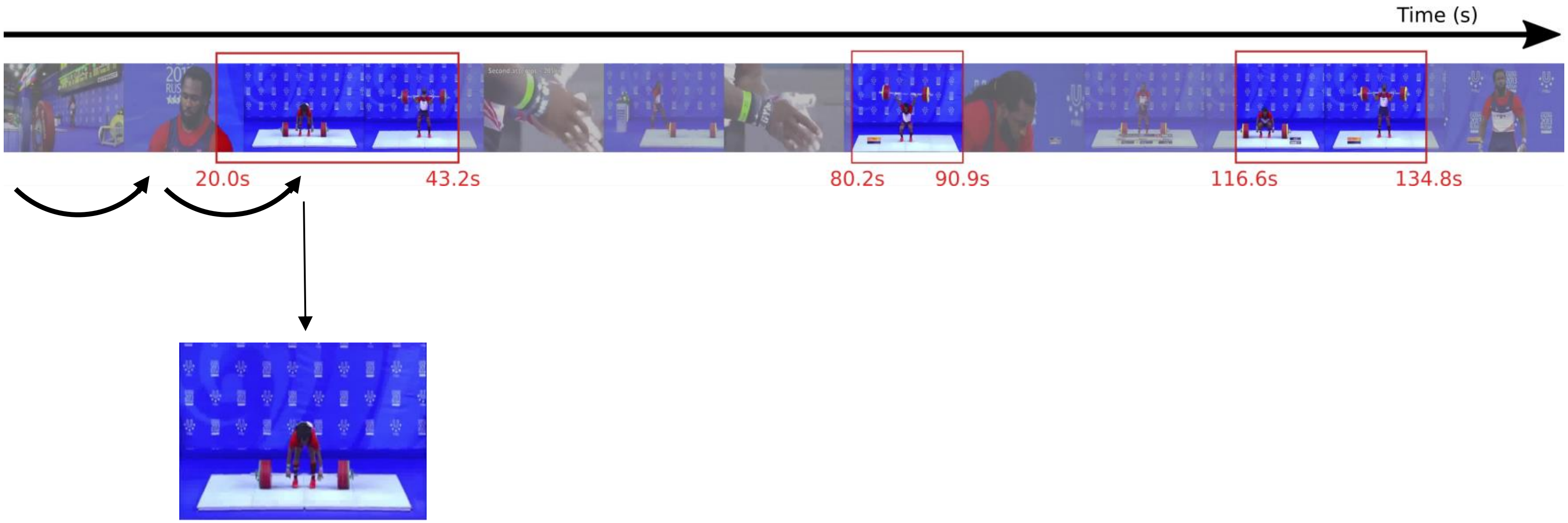
Action Spotting



t = 26.4s

Problem setup

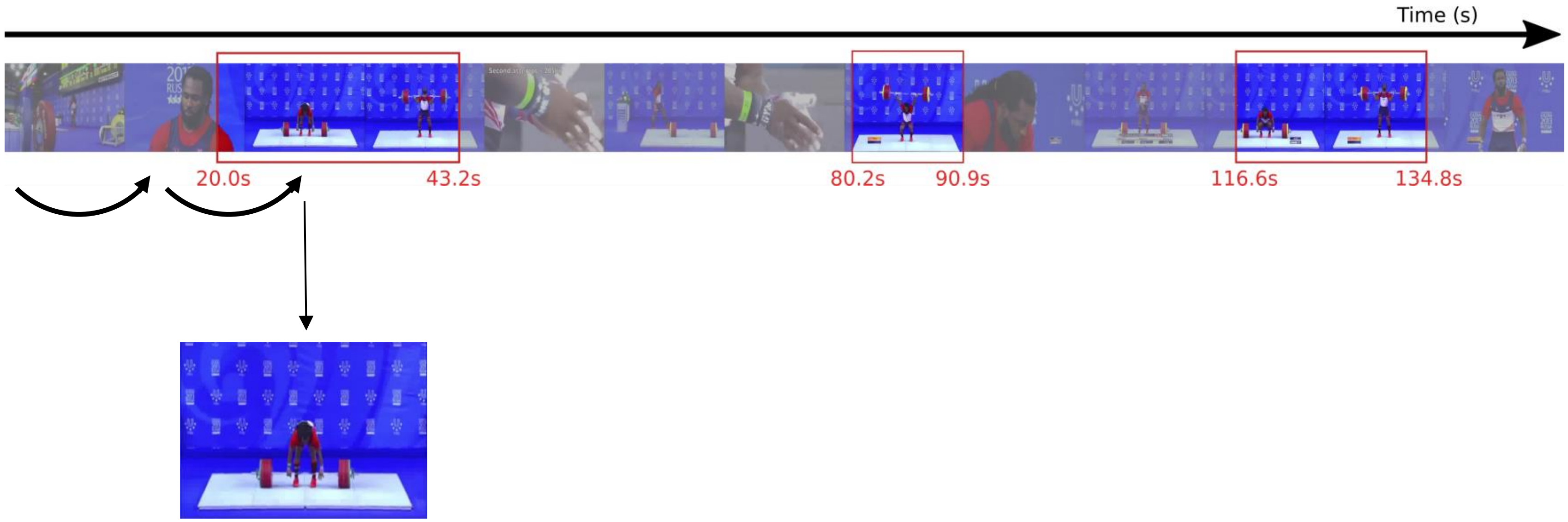
Action Spotting



t = 26.4s
p = 0.59

Problem setup

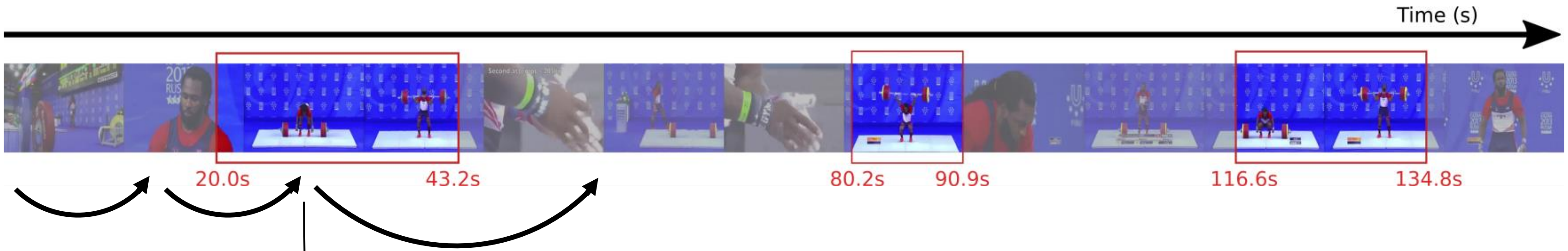
Action Spotting



t = 26.4s
p = 0.59
Clean & Jerk

Problem setup

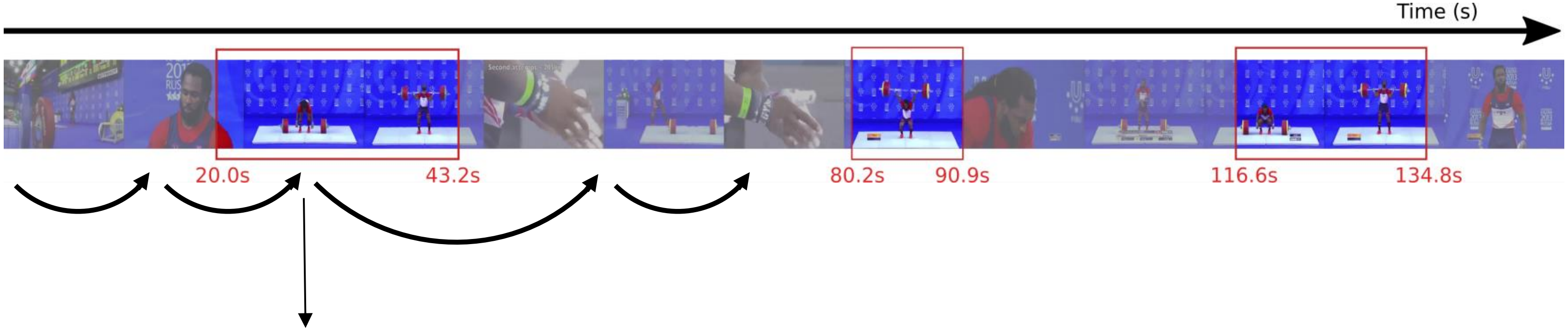
Action Spotting



t = 26.4s
p = 0.59
Clean & Jerk

Problem setup

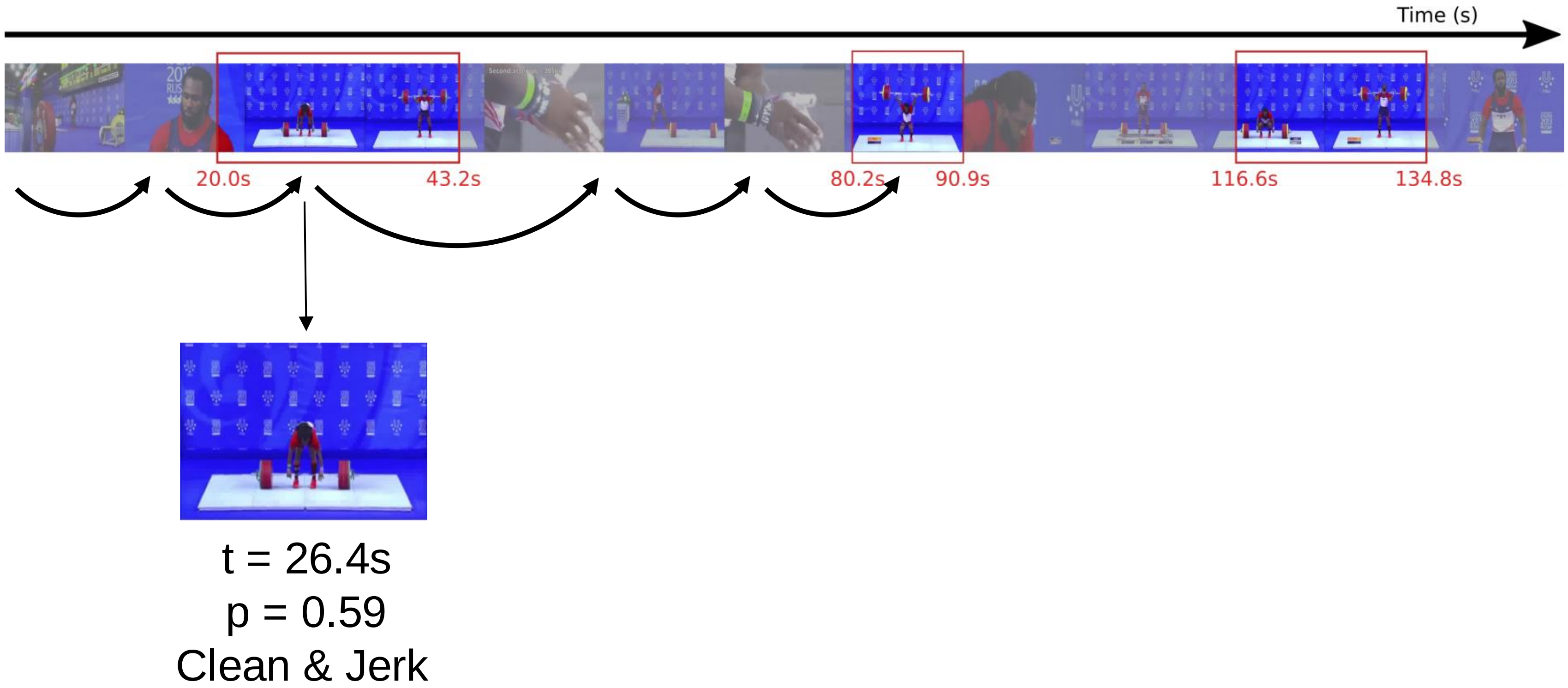
Action Spotting



t = 26.4s
p = 0.59
Clean & Jerk

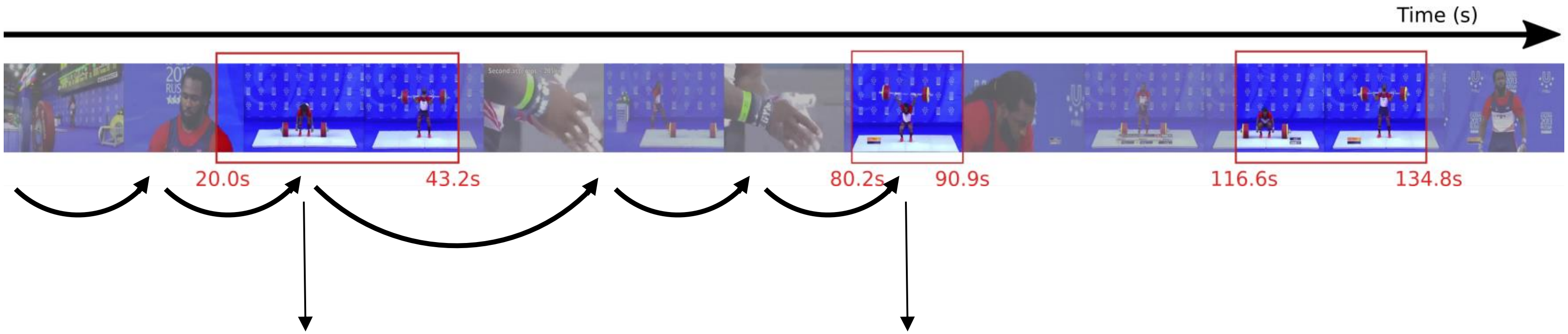
Problem setup

Action Spotting



Problem setup

Action Spotting

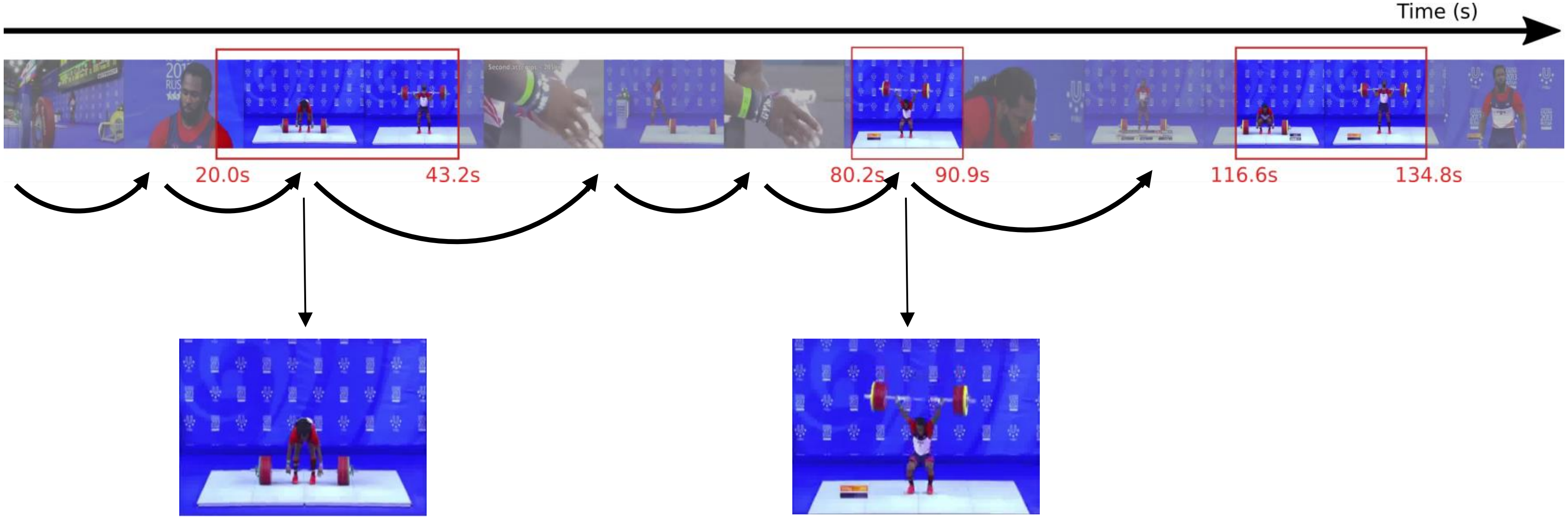


t = 26.4s
p = 0.59
Clean & Jerk

t = 83.7s
p = 0.96
Clean & Jerk

Problem setup

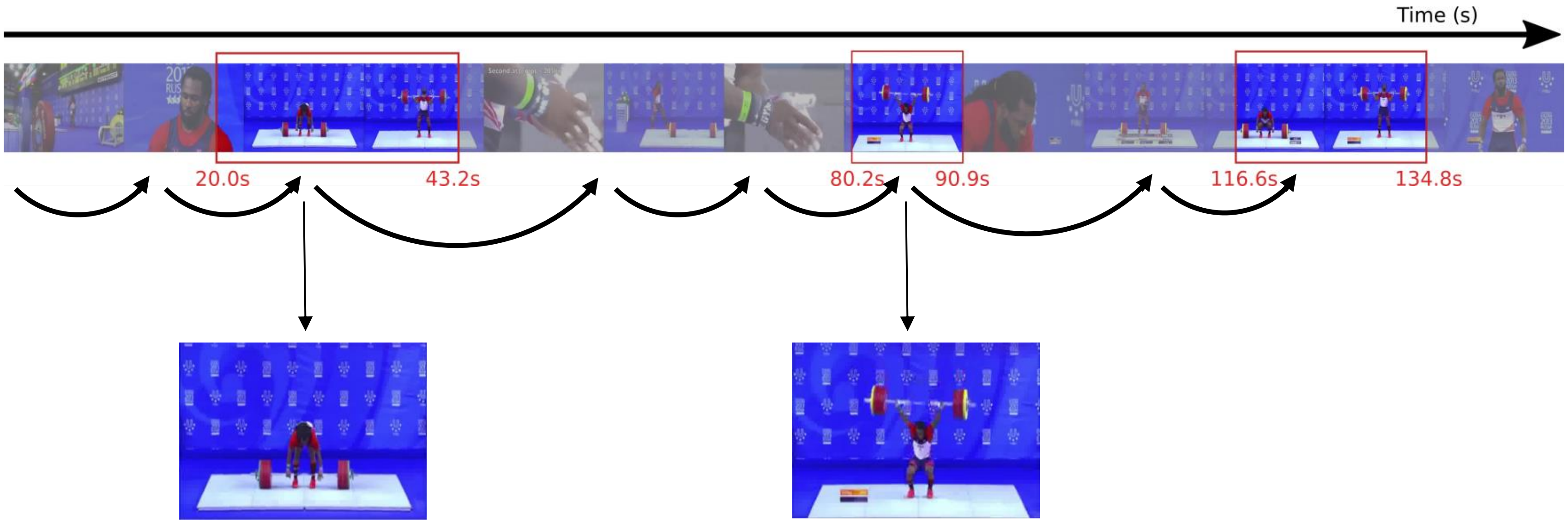
Action Spotting



t = 26.4s
p = 0.59
Clean & Jerk

t = 83.7s
p = 0.96
Clean & Jerk

Action Spotting

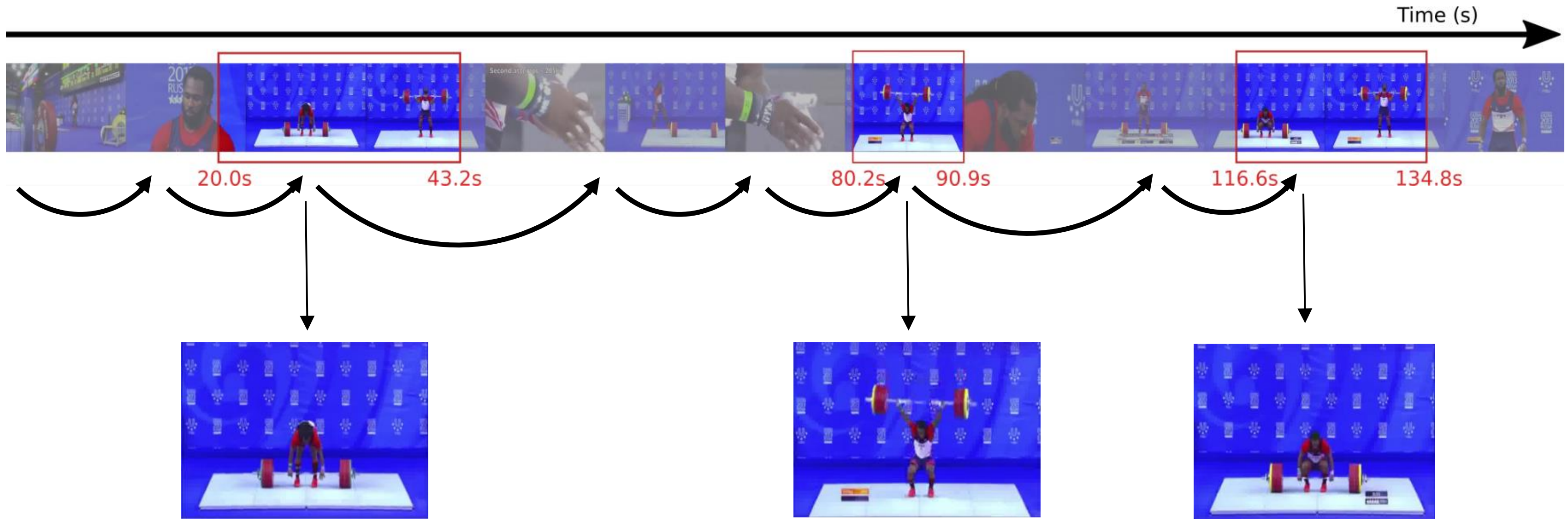


t = 26.4s
p = 0.59
Clean & Jerk

t = 83.7s
p = 0.96
Clean & Jerk

Problem setup

Action Spotting



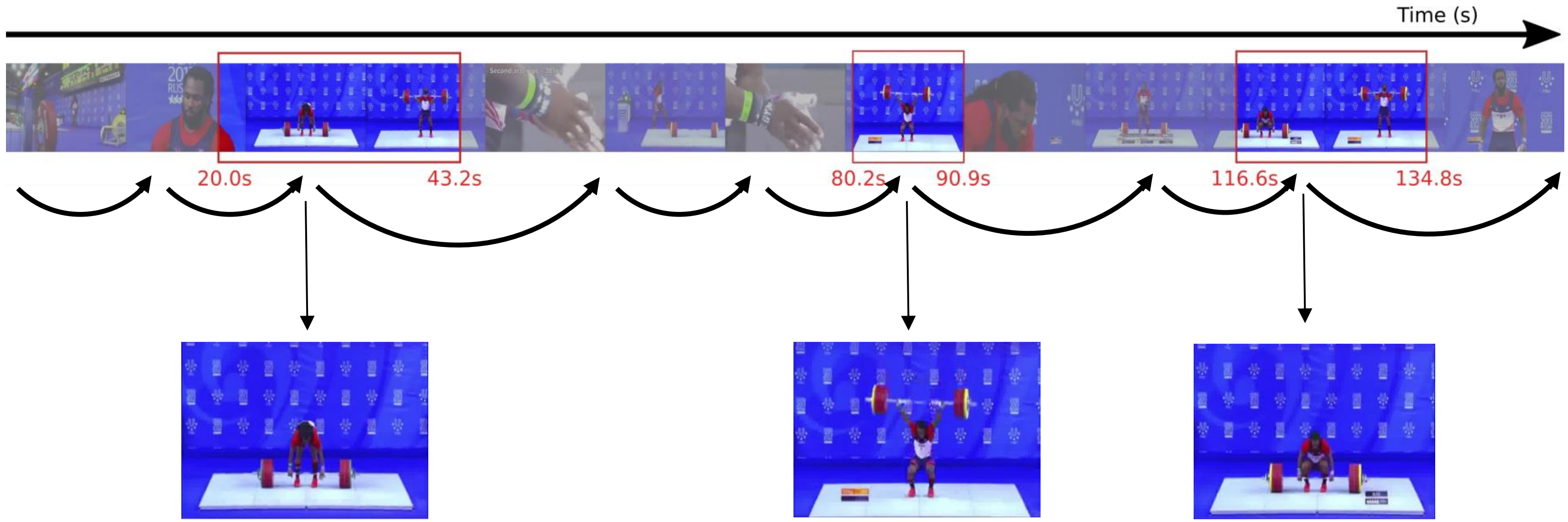
t = 26.4s
p = 0.59
Clean & Jerk

t = 83.7s
p = 0.96
Clean & Jerk

t = 120.1s
p = 0.45
Clean & Jerk

Problem setup

Action Spotting

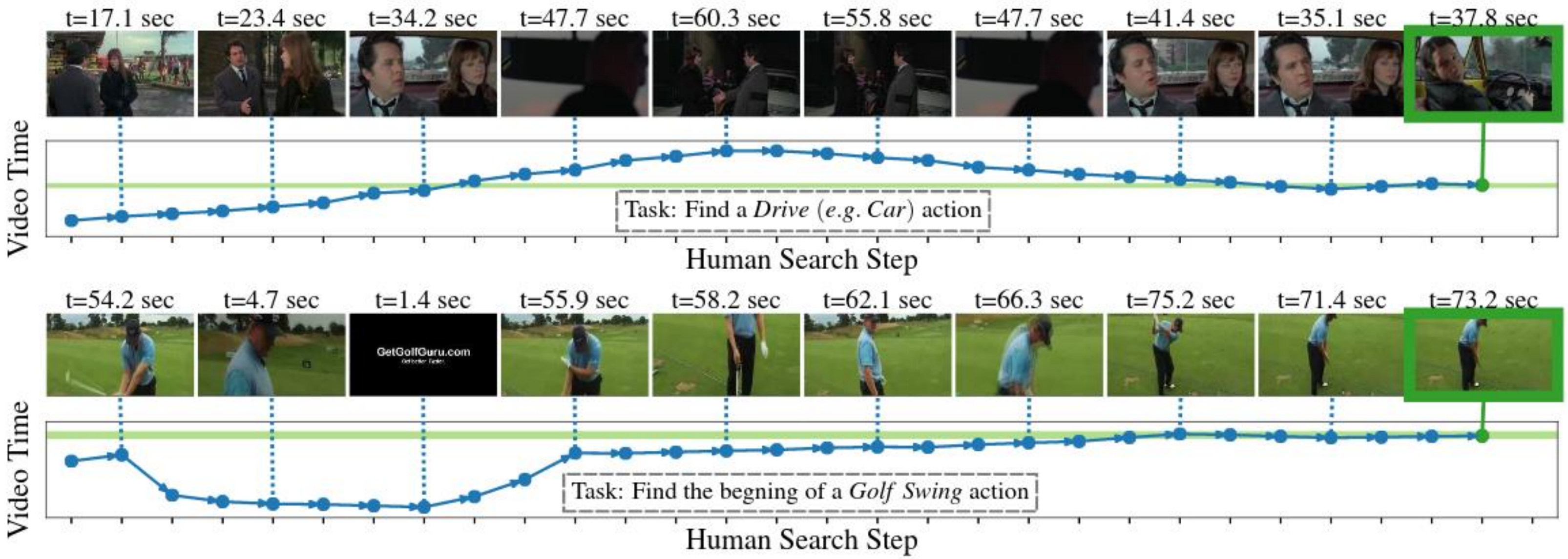


t = 26.4s
p = 0.59
Clean & Jerk

t = 83.7s
p = 0.96
Clean & Jerk

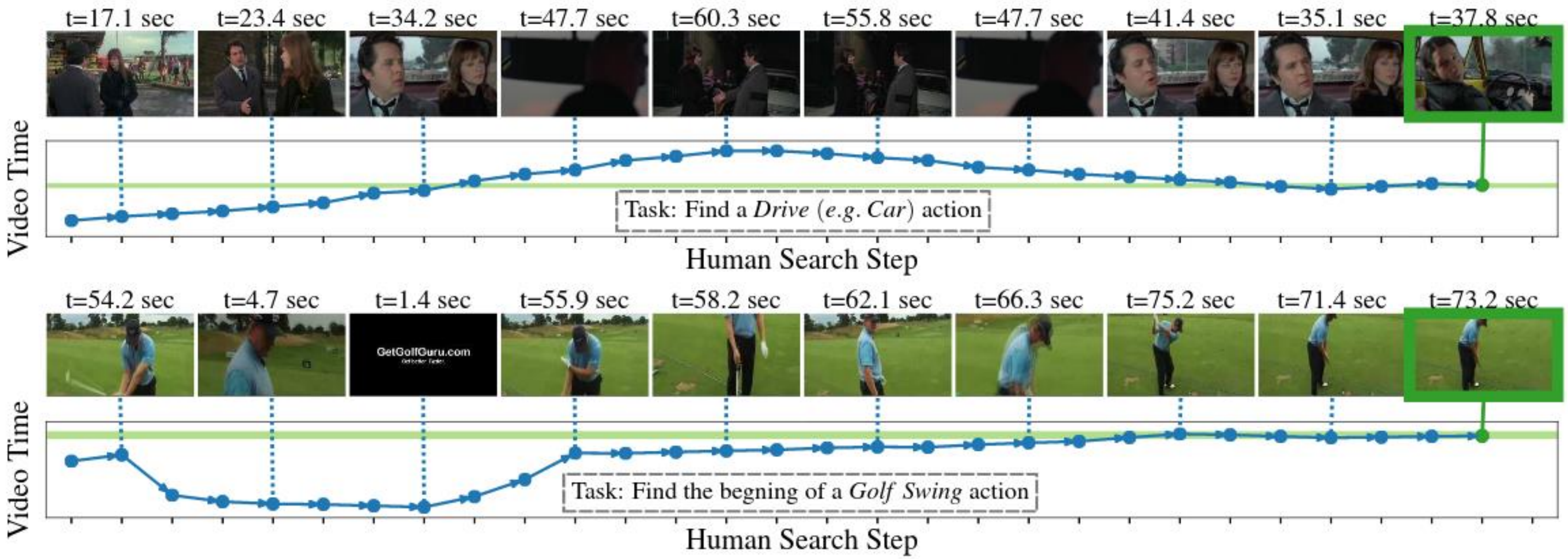
t = 120.1s
p = 0.45
Clean & Jerk

State-of-the-Art



Alwassel et al., ECCV 2018

State-of-the-Art



Alwassel et al., ECCV 2018

Requires a lot of human acquisitions

ActionSpotter: Approach overview

Asymmetrical problem

ActionSpotter: Approach overview

Asymmetrical problem

Use of reinforcement learning

ActionSpotter: Approach overview

Asymmetrical problem

Use of reinforcement learning

Only need action detection ground-truth

ActionSpotter: Approach overview

Asymmetrical problem

- Use of reinforcement learning
- Only need action detection ground-truth
- Extracting spot frames while observing as few frames as possible

ActionSpotter: Approach overview

Asymmetrical problem

- Use of reinforcement learning
- Only need action detection ground-truth
- Extracting spot frames while observing as few frames as possible
- End-to-End training

ActionSpotter: Approach overview

Asymmetrical problem

Use of reinforcement learning

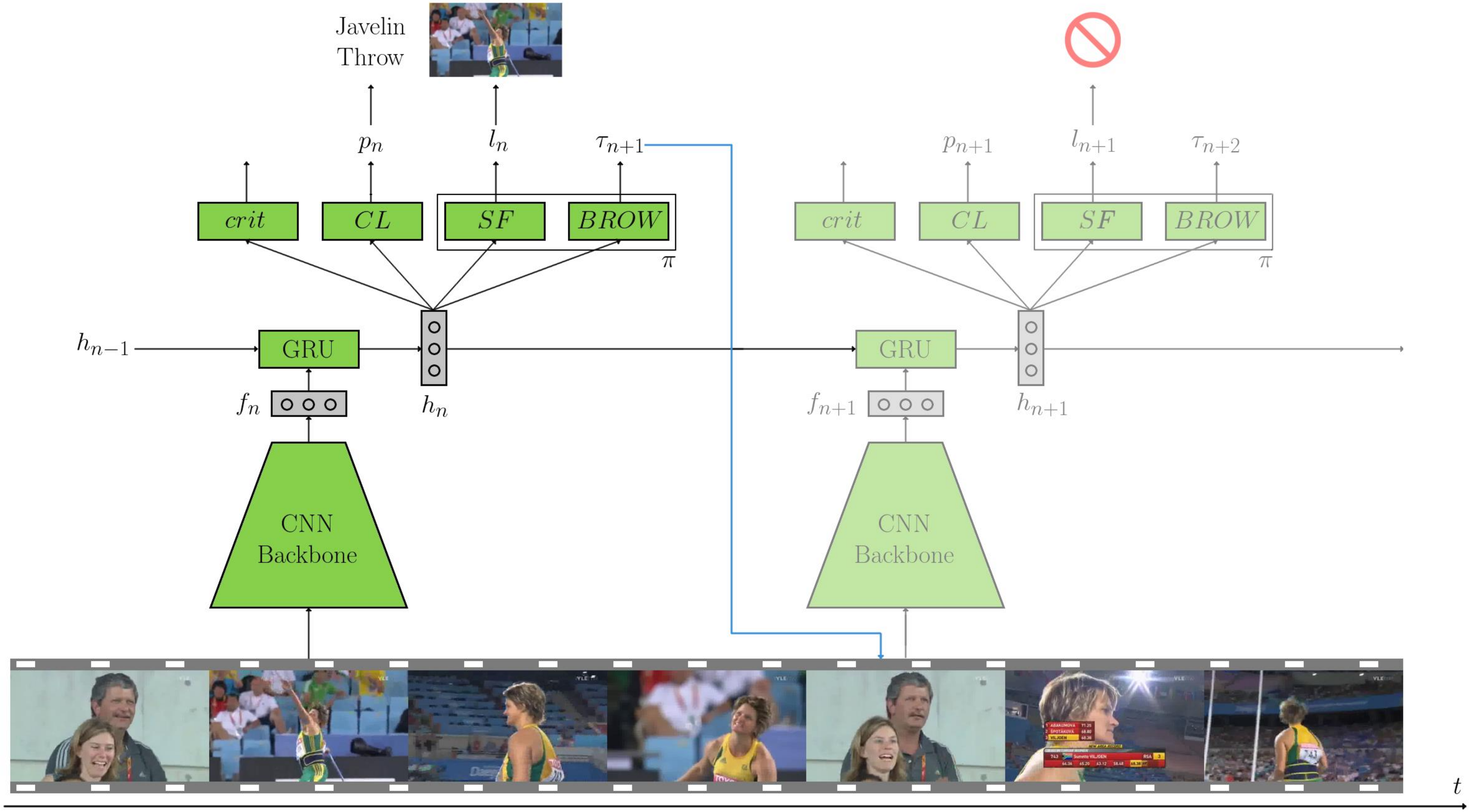
Only need action detection ground-truth

Extracting spot frames while observing as few frames as possible

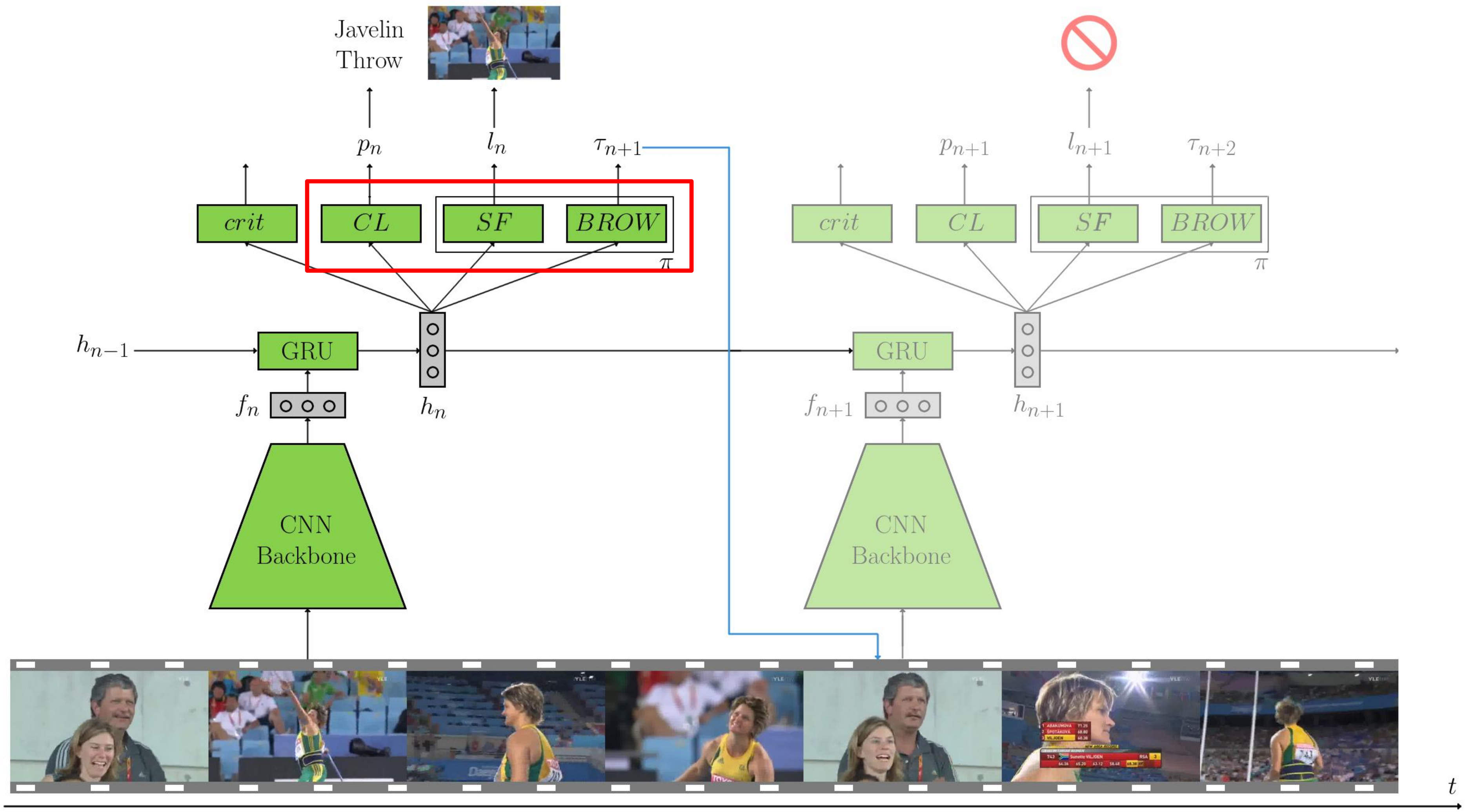
End-to-End training

Increase efficiency on action spotting task

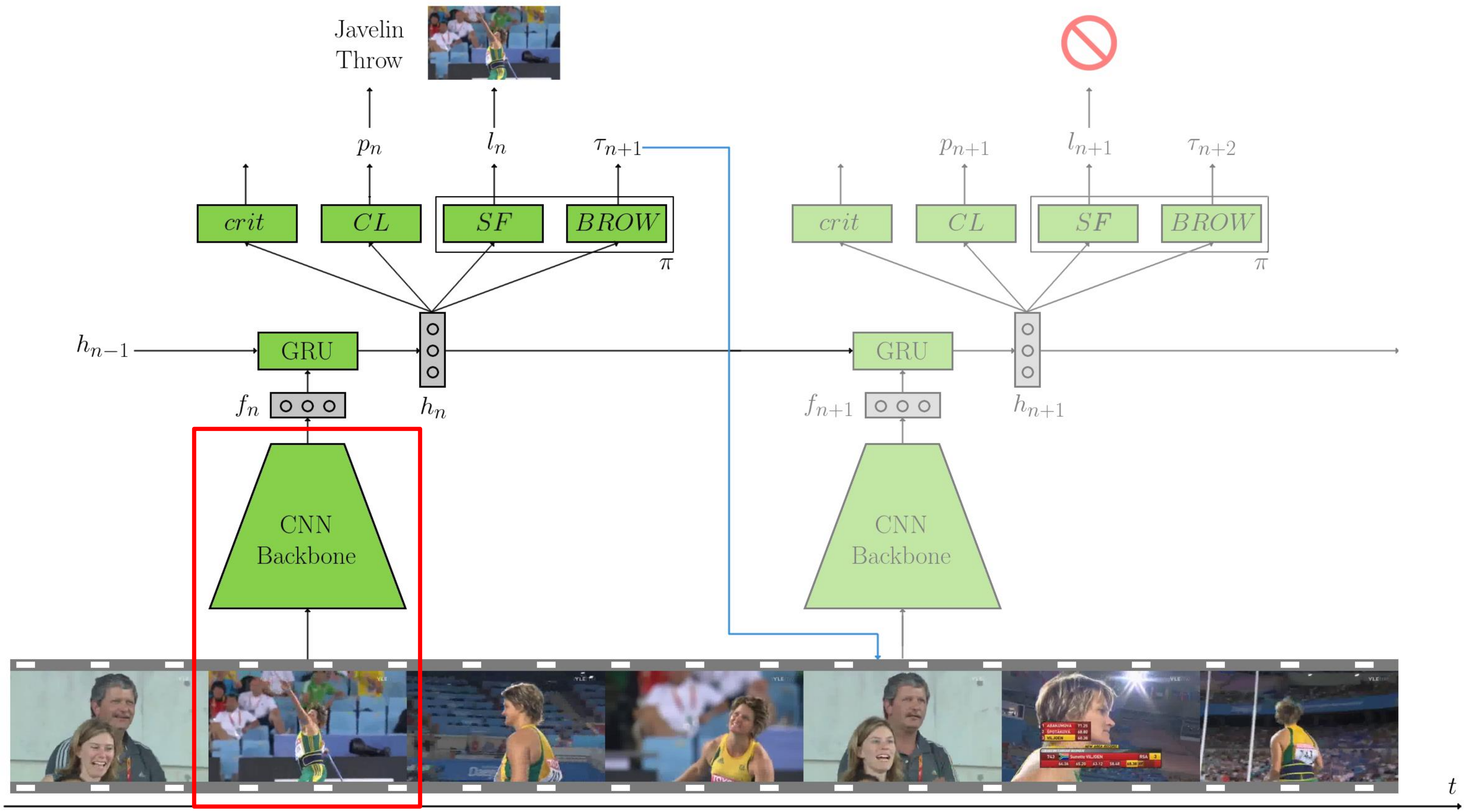
ActionSpotter: RL Framework



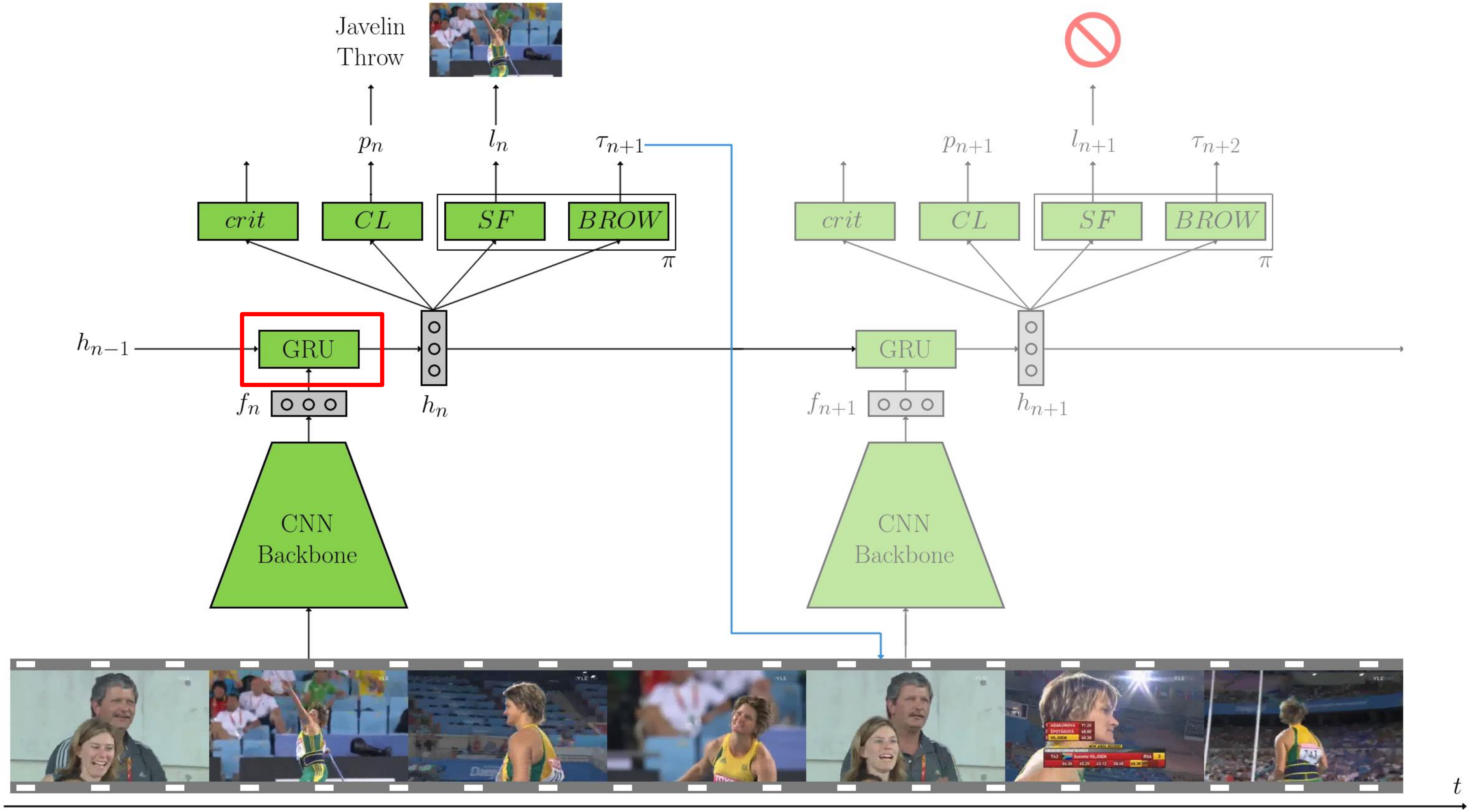
ActionSpotter: RL Framework



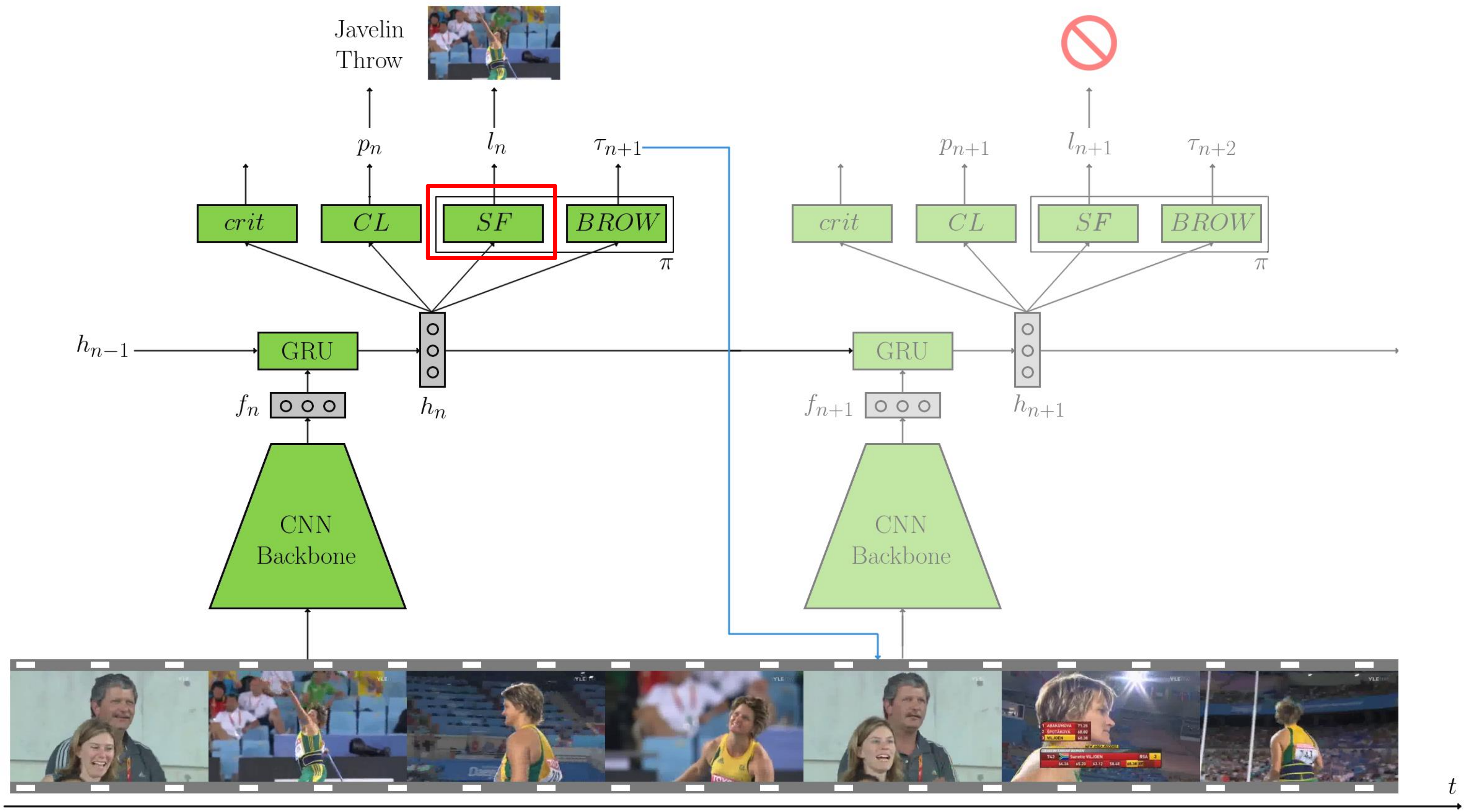
ActionSpotter: RL Framework



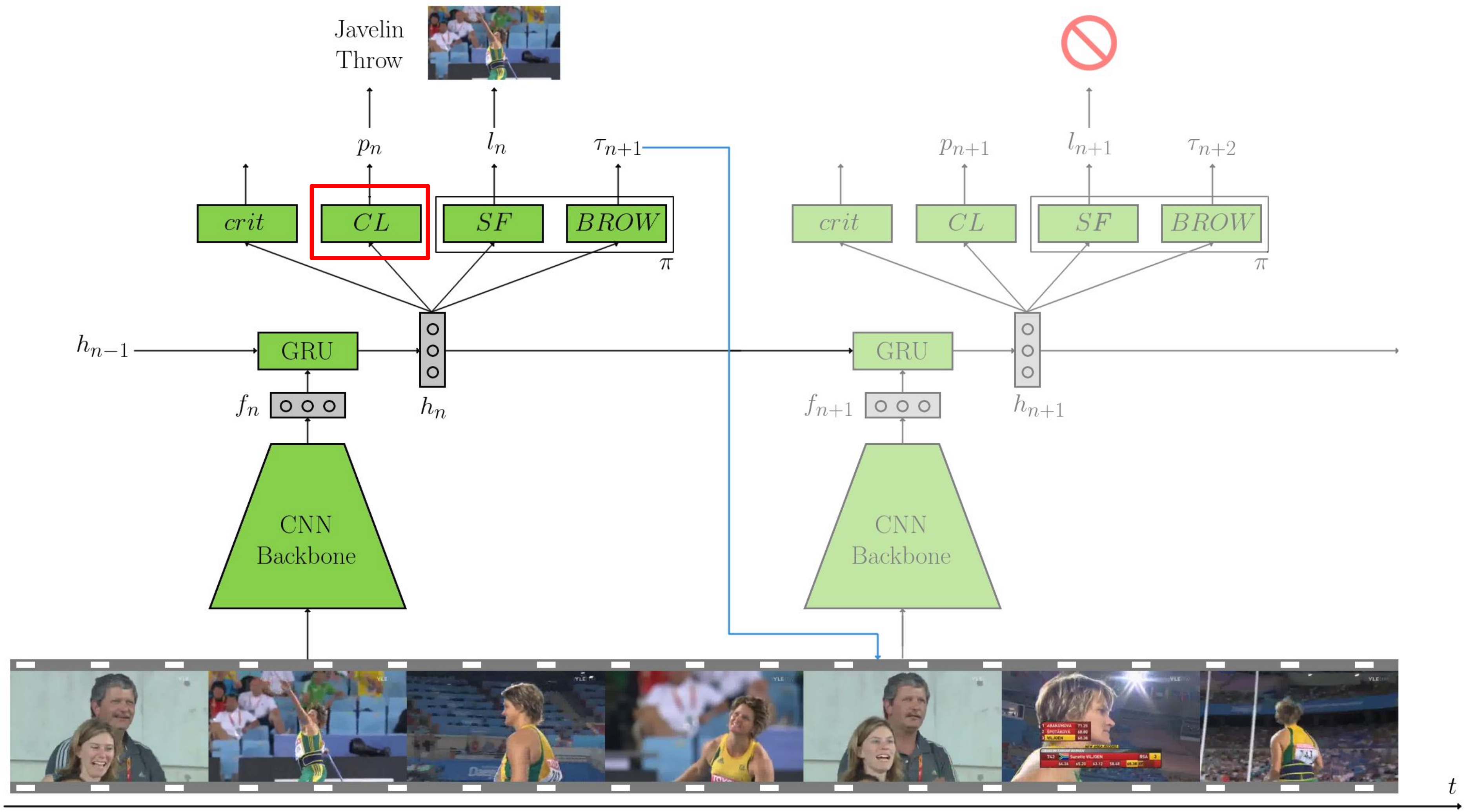
ActionSpotter: RL Framework



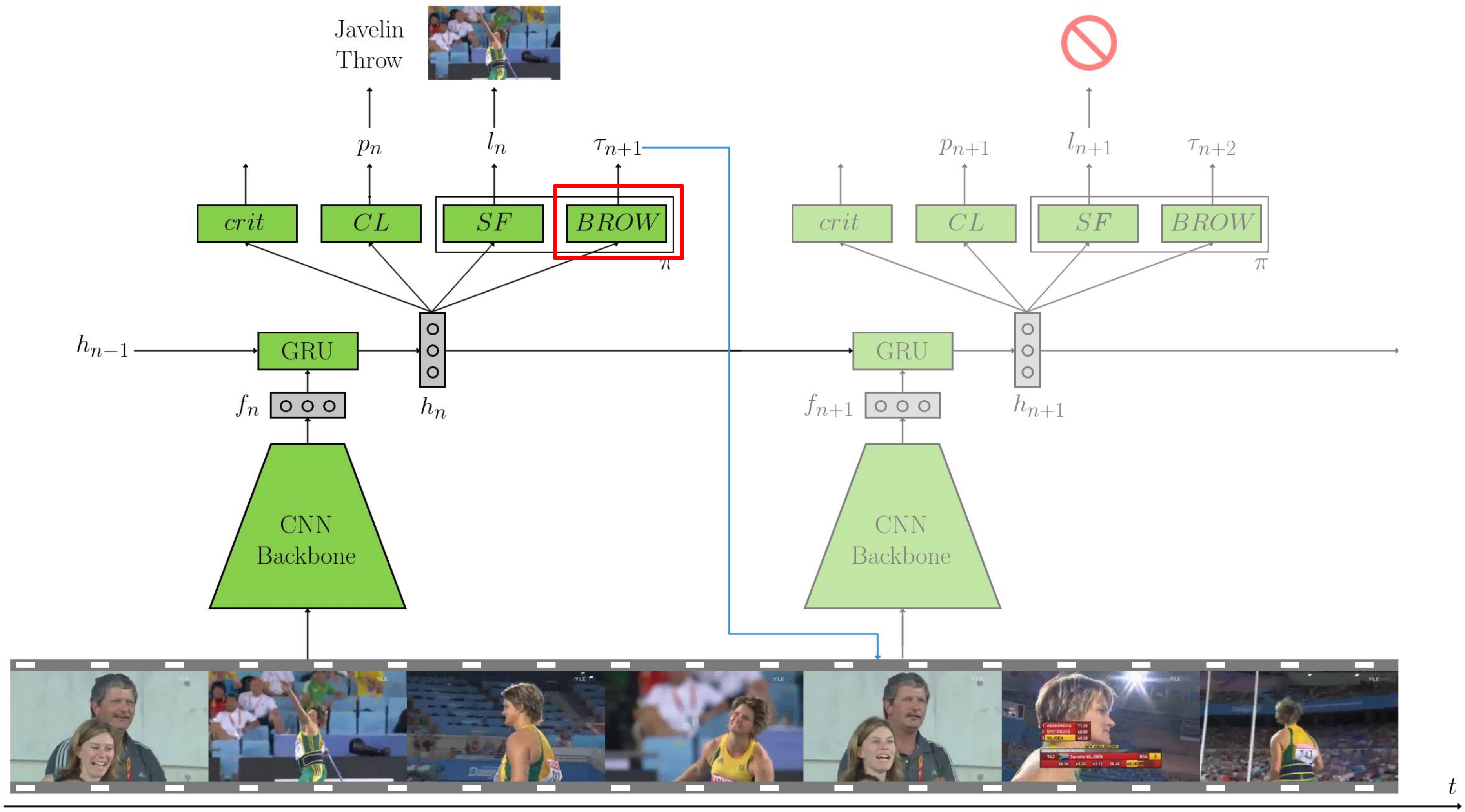
ActionSpotter: RL Framework



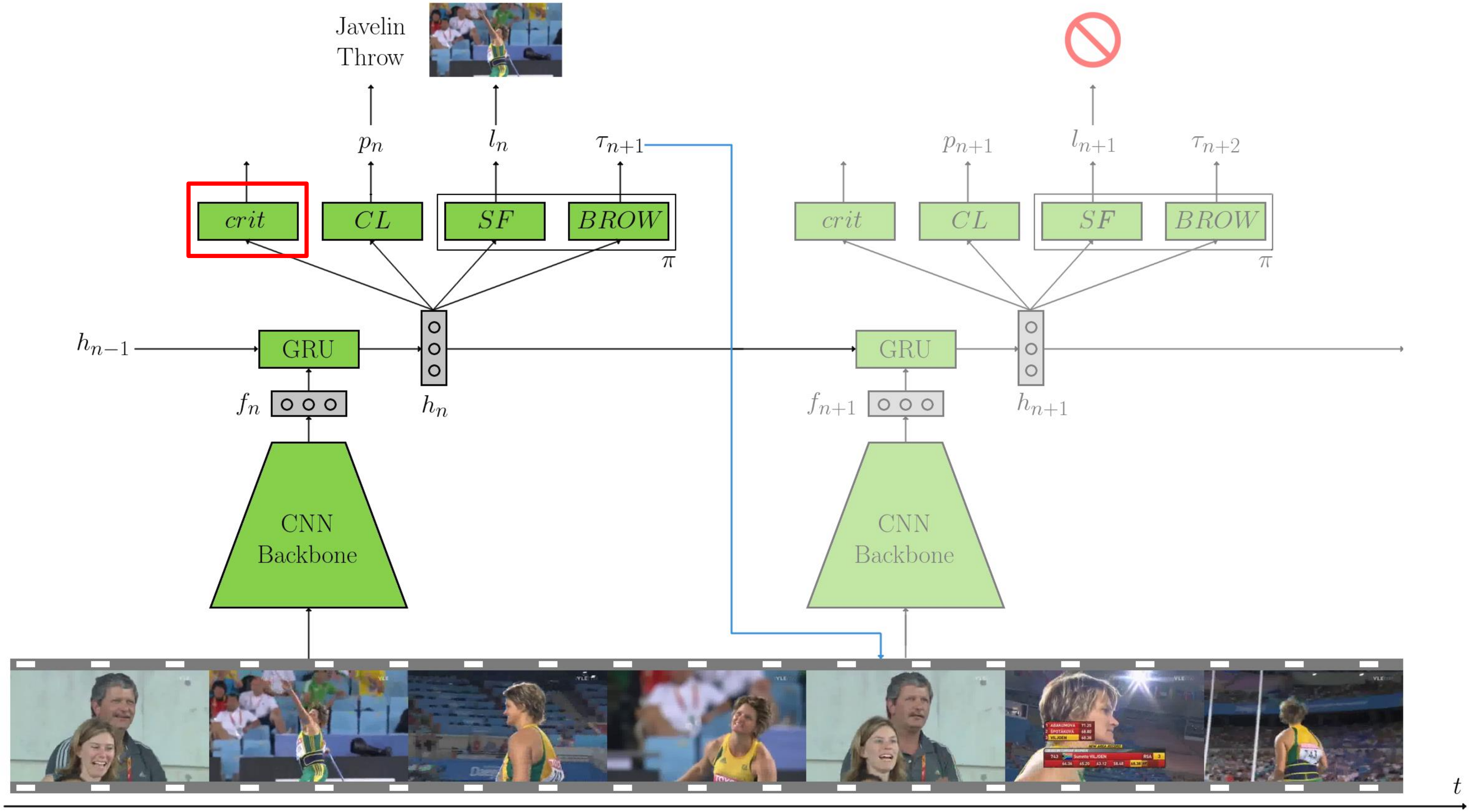
ActionSpotter: RL Framework



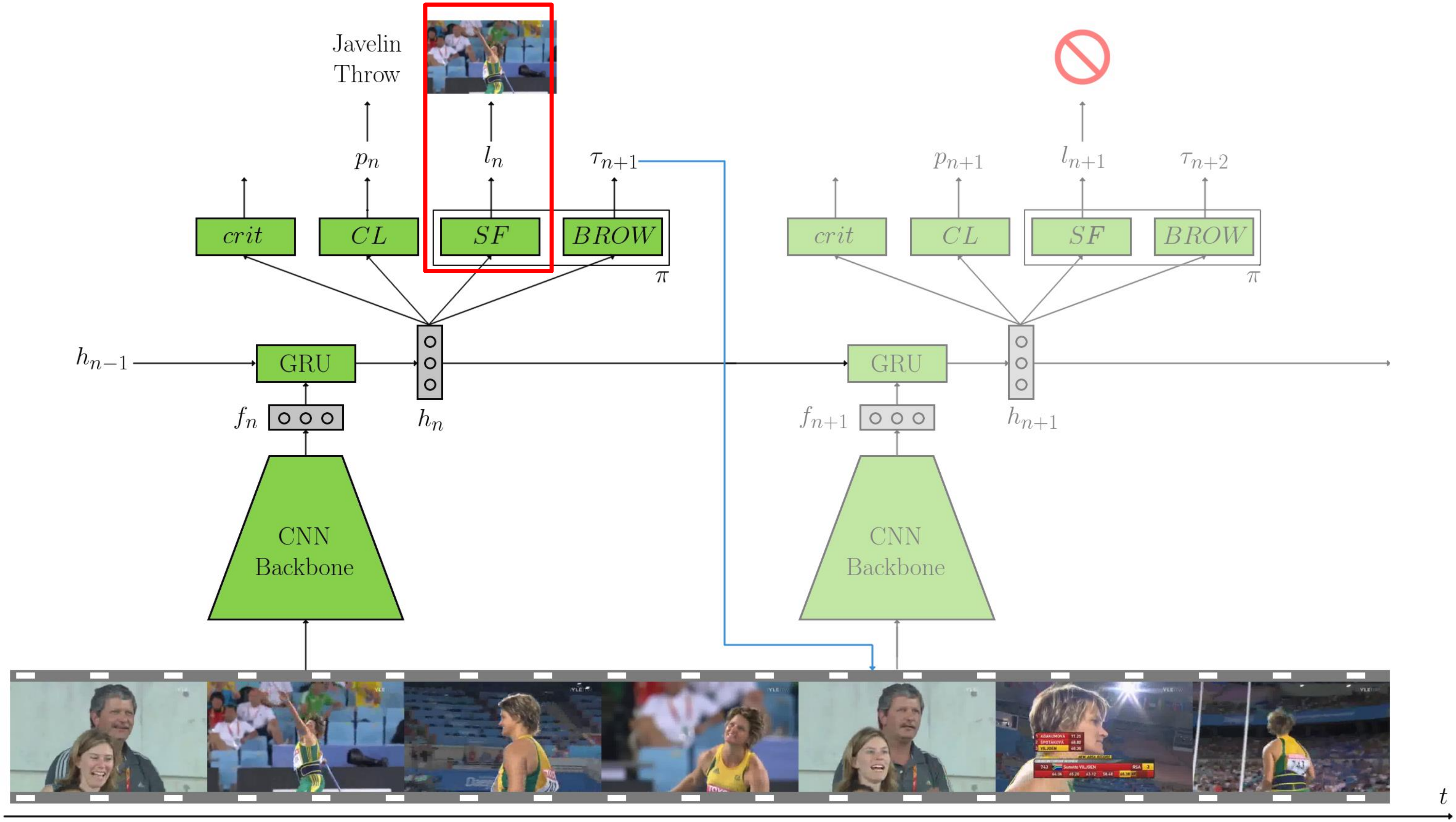
ActionSpotter: RL Framework



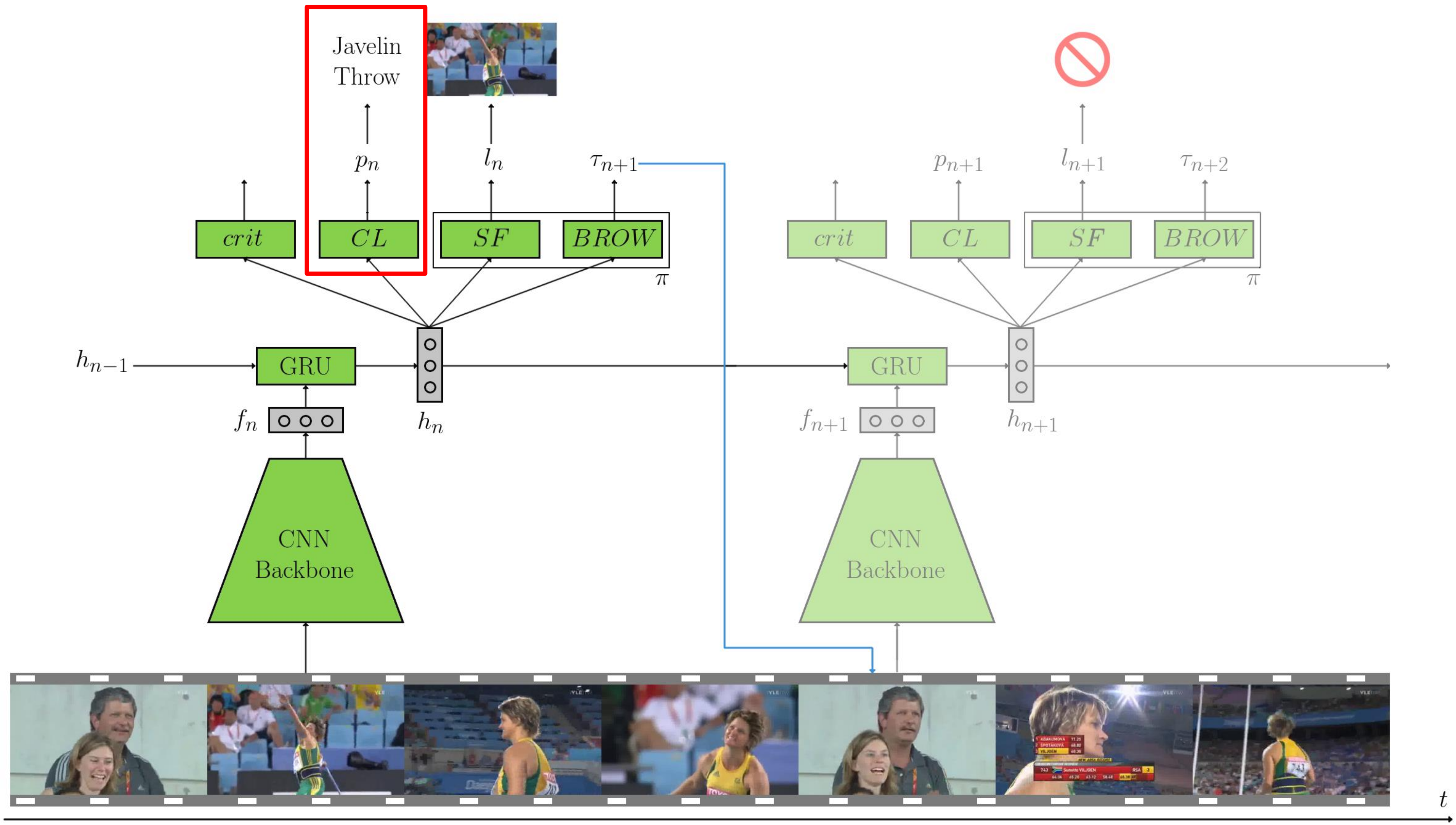
ActionSpotter: RL Framework



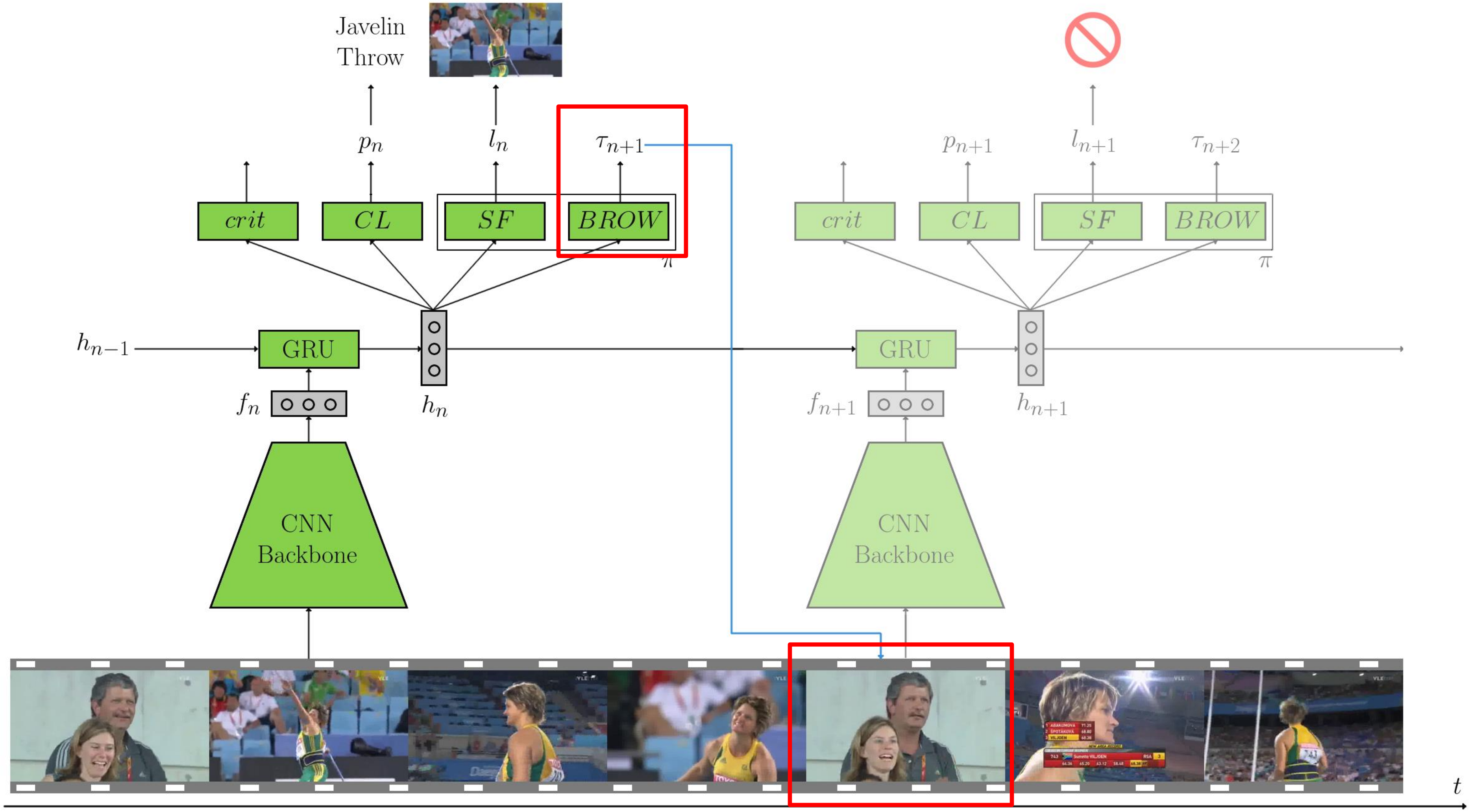
ActionSpotter: RL Framework



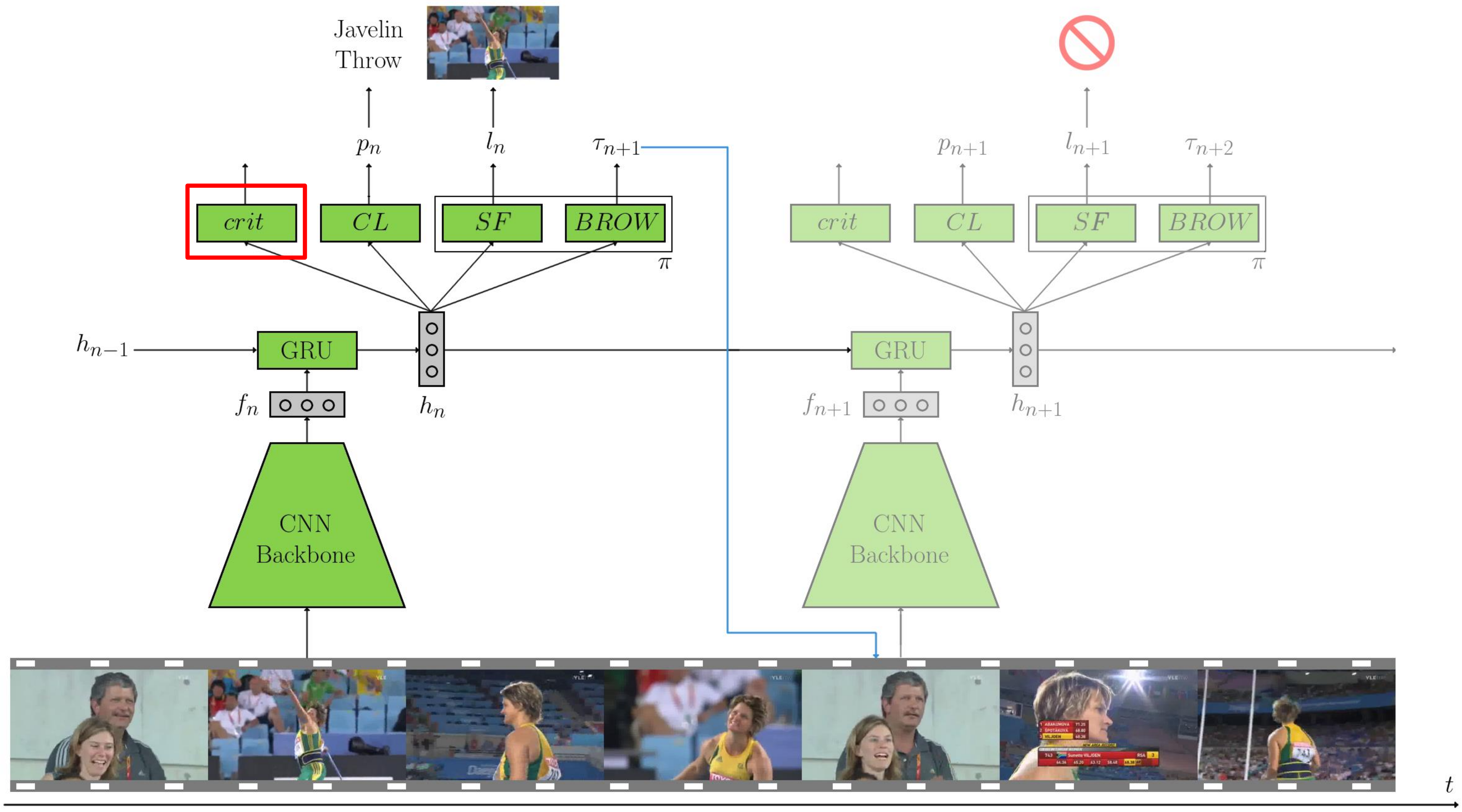
ActionSpotter: RL Framework



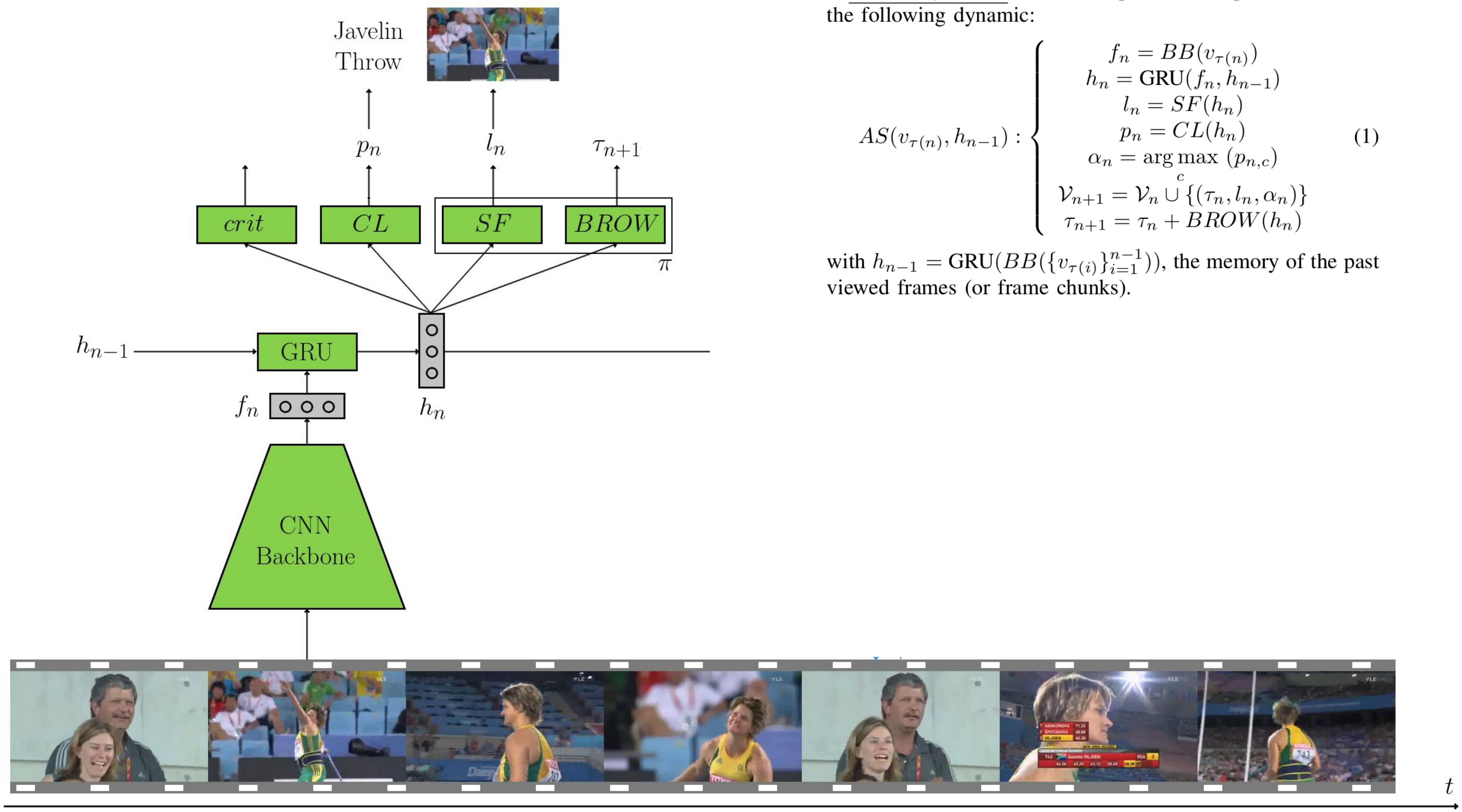
ActionSpotter: RL Framework



ActionSpotter: RL Framework



ActionSpotter: RL Framework

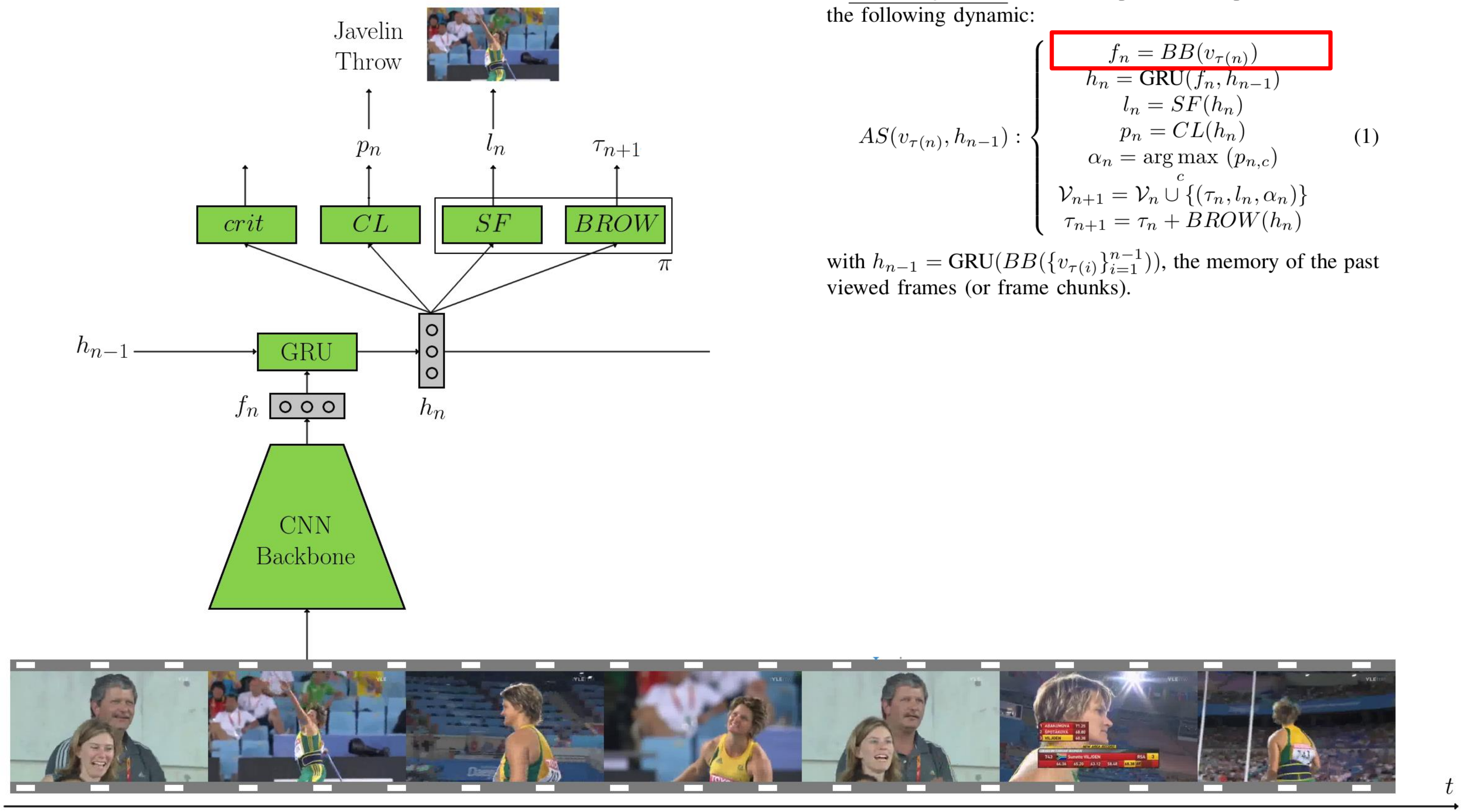


Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c\,(p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

ActionSpotter: RL Framework

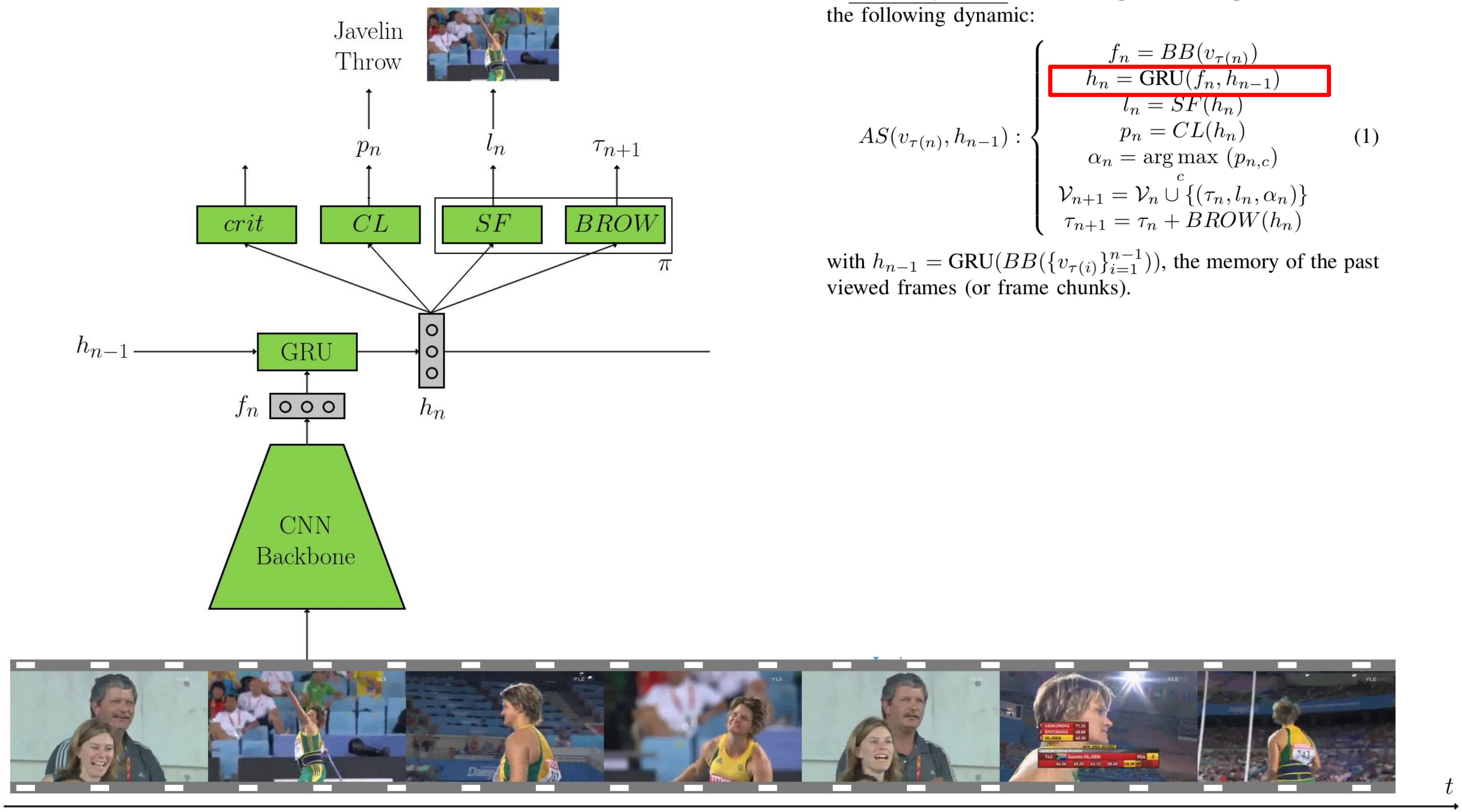


Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} \boxed{f_n = BB(v_{\tau(n)})} \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c \, (p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

ActionSpotter: RL Framework

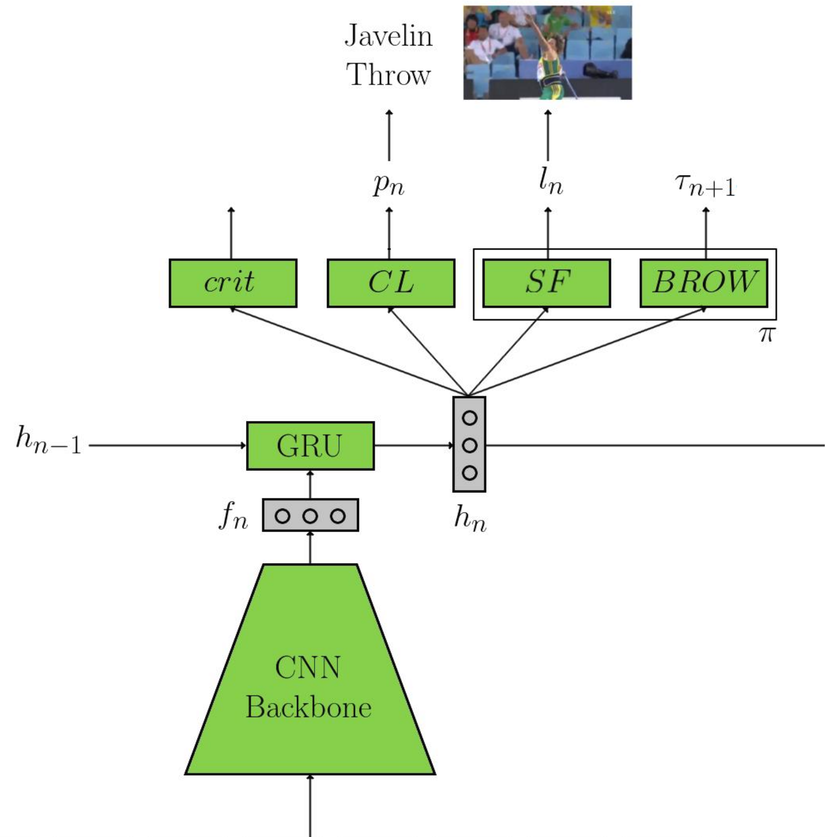


Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ \boxed{h_n = \text{GRU}(f_n, h_{n-1})} \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c\,(p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

ActionSpotter: RL Framework

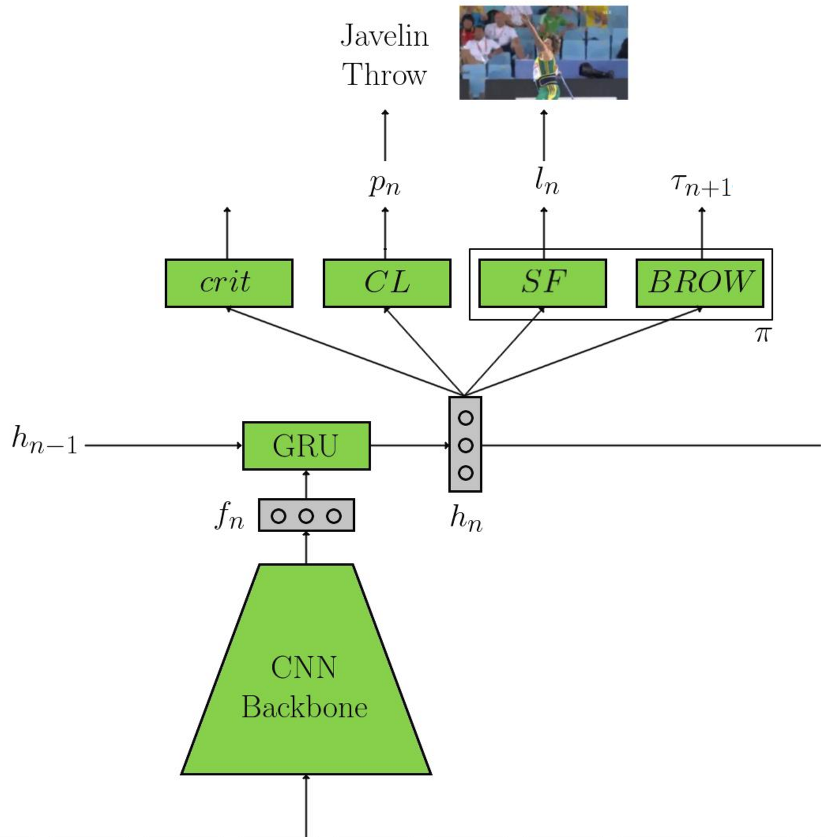


Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \left\{ \begin{aligned} f_n &= BB(v_{\tau(n)}) \\ h_n &= \text{GRU}(f_n, h_{n-1}) \\ &\boxed{\begin{aligned} l_n &= SF(h_n) \\ p_n &= CL(h_n) \\ \alpha_n &= \arg\max\,(p_{n,c}) \end{aligned}} \\ \mathcal{V}_{n+1} &= \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} &= \tau_n + BROW(h_n) \end{aligned} \right. \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

ActionSpotter: RL Framework

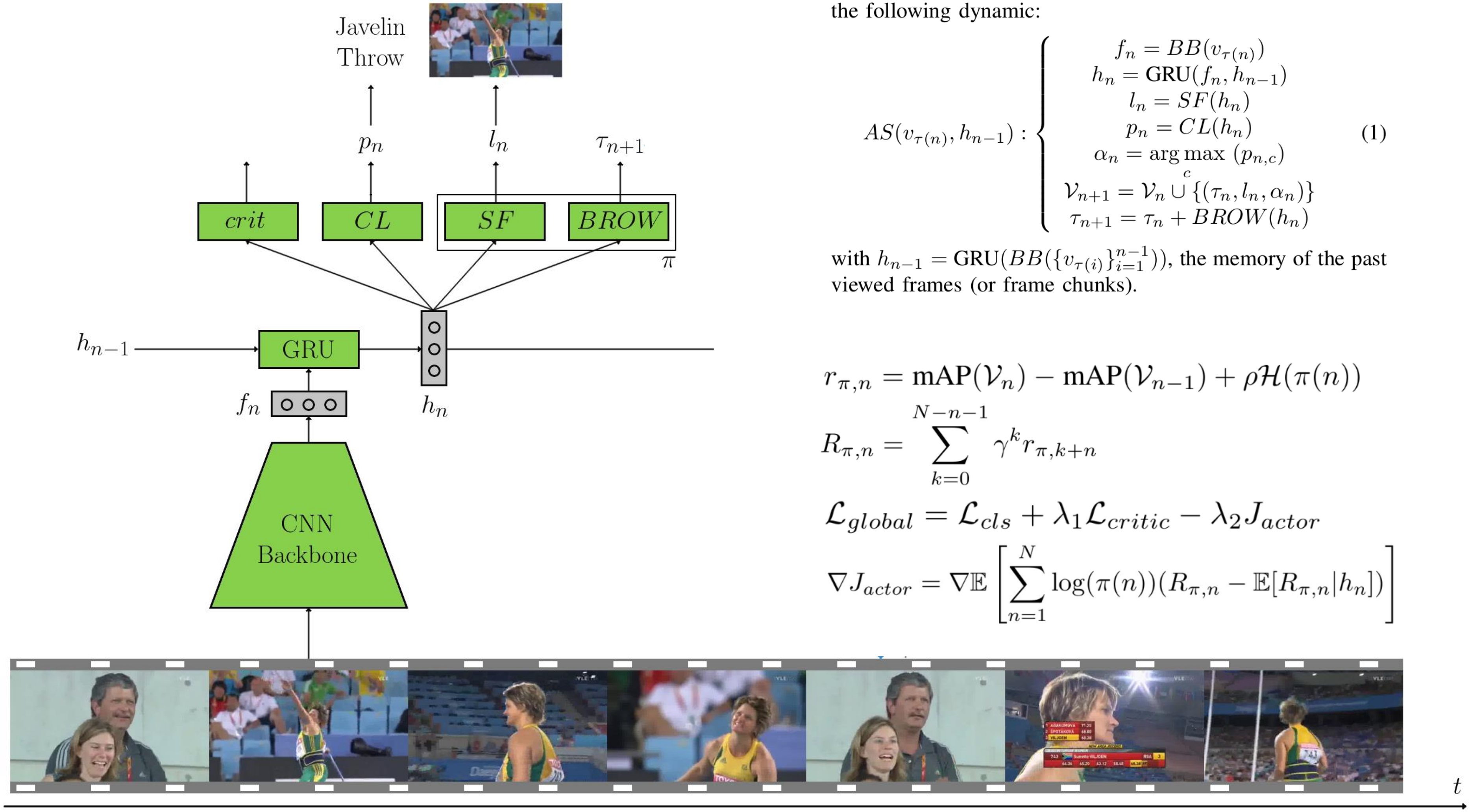


Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c (p_{n,c}) \\ \boxed{\begin{array}{c} \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{array}} \end{cases} \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

ActionSpotter: RL Framework



Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c \, (p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

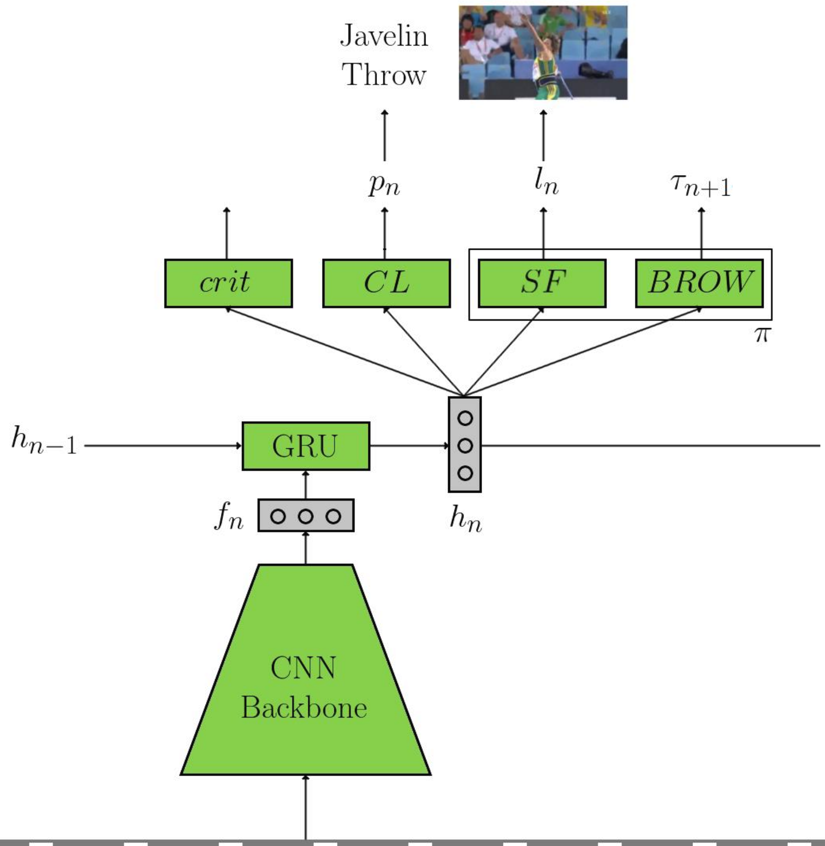
$$r_{\pi,n} = \text{mAP}(\mathcal{V}_n) - \text{mAP}(\mathcal{V}_{n-1}) + \rho\mathcal{H}(\pi(n))$$

$$R_{\pi,n} = \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n}$$

$$\mathcal{L}_{global} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor}$$

$$\nabla J_{actor} = \nabla \mathbb{E}\left[\sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n])\right]$$

ActionSpotter: RL Framework



Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c\ (p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \qquad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

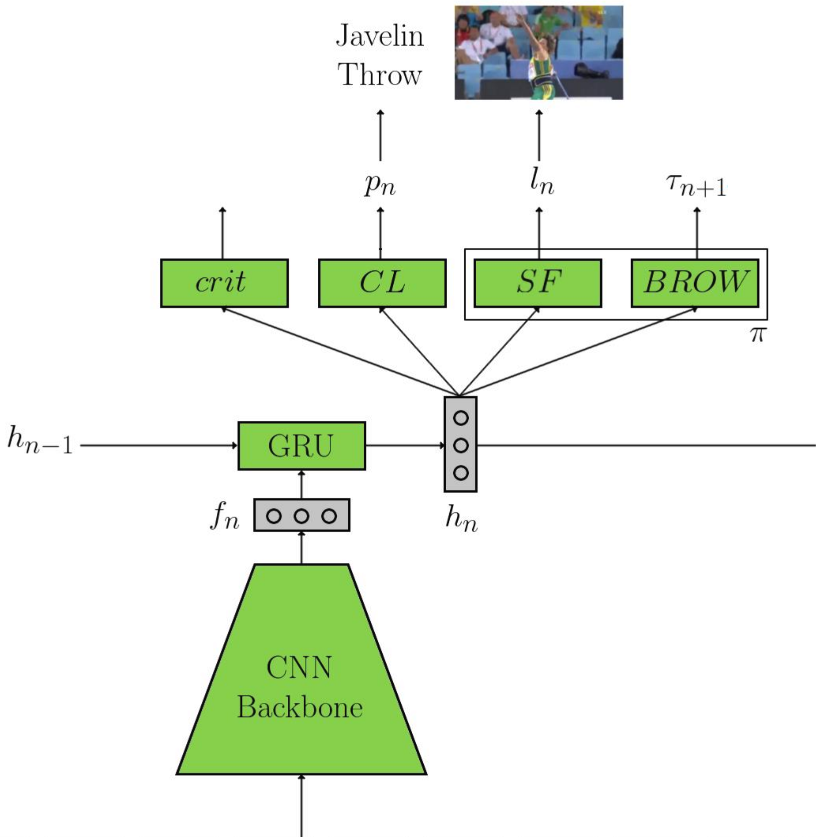
$$\boxed{r_{\pi,n} = \text{mAP}(\mathcal{V}_n) - \text{mAP}(\mathcal{V}_{n-1}) + \rho\mathcal{H}(\pi(n))}$$

$$R_{\pi,n} = \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n}$$

$$\mathcal{L}_{global} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor}$$

$$\nabla J_{actor} = \nabla \mathbb{E}\left[\sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n])\right]$$

ActionSpotter: RL Framework



Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c \, (p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

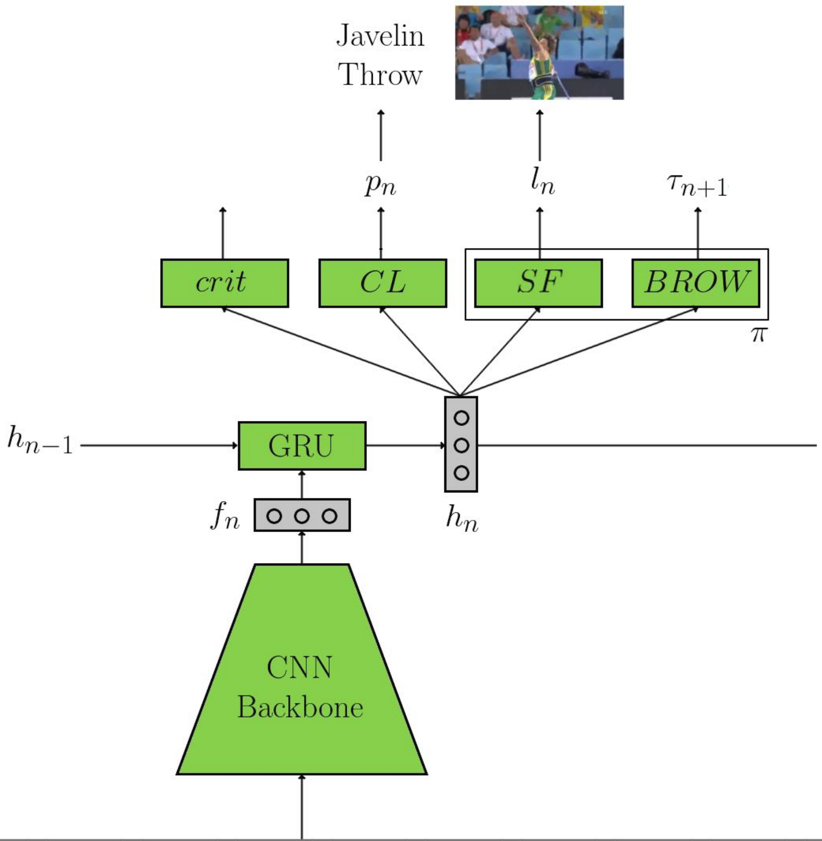
$$r_{\pi,n} = \text{mAP}(\mathcal{V}_n) - \text{mAP}(\mathcal{V}_{n-1}) + \rho\mathcal{H}(\pi(n))$$

$$\boxed{R_{\pi,n} = \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n}}$$

$$\mathcal{L}_{global} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor}$$

$$\nabla J_{actor} = \nabla \mathbb{E}\left[\sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n])\right]$$

ActionSpotter: RL Framework



Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c \, (p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \tag{1}$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

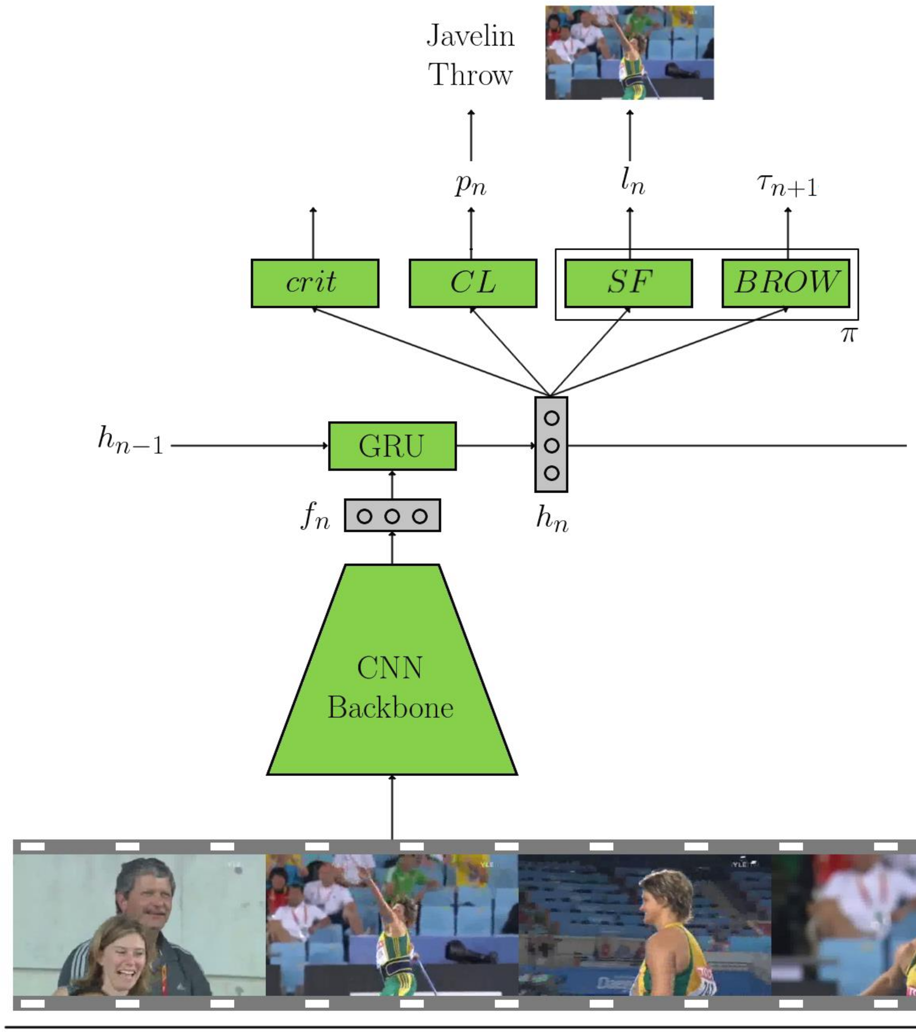
$$r_{\pi,n} = \text{mAP}(\mathcal{V}_n) - \text{mAP}(\mathcal{V}_{n-1}) + \rho\mathcal{H}(\pi(n))$$

$$R_{\pi,n} = \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n}$$

$$\boxed{\mathcal{L}_{global} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor}}$$

$$\nabla J_{actor} = \nabla \mathbb{E}\left[\sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n])\right]$$

ActionSpotter: RL Framework



Global dynamic: Then, at step n, ActionSpotter (AS), has the following dynamic:

$$AS(v_{\tau(n)}, h_{n-1}) : \begin{cases} f_n = BB(v_{\tau(n)}) \\ h_n = \text{GRU}(f_n, h_{n-1}) \\ l_n = SF(h_n) \\ p_n = CL(h_n) \\ \alpha_n = \arg\max_c (p_{n,c}) \\ \mathcal{V}_{n+1} = \mathcal{V}_n \cup \{(\tau_n, l_n, \alpha_n)\} \\ \tau_{n+1} = \tau_n + BROW(h_n) \end{cases} \quad (1)$$

with $h_{n-1} = \text{GRU}(BB(\{v_{\tau(i)}\}_{i=1}^{n-1}))$, the memory of the past viewed frames (or frame chunks).

$$r_{\pi,n} = \text{mAP}(\mathcal{V}_n) - \text{mAP}(\mathcal{V}_{n-1}) + \rho\mathcal{H}(\pi(n))$$

$$R_{\pi,n} = \sum_{k=0}^{N-n-1} \gamma^k r_{\pi,k+n}$$

$$\mathcal{L}_{global} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{critic} - \lambda_2 J_{actor}$$

$$\boxed{\nabla J_{actor} = \nabla\mathbb{E}\left[\sum_{n=1}^{N} \log(\pi(n))(R_{\pi,n} - \mathbb{E}[R_{\pi,n}|h_n])\right]}$$

Experiments

- Thumos14

THUMOS'14						
Approach	Detection mAP@					Spotting mAP
	0.1	0.2	0.3	0.4	0.5	
Glimpses [5]	48.9	44.0	36.0	26.4	17.1	-
SMS [30]	51.0	45.2	36.5	27.8	17.8	-
M-CNN [31]	47.7	43.5	36.3	28.7	19.0	41.2
CDC [32]	-	-	41.3	30.7	24.7	31.5
TURN [33]	54.0	50.9	44.1	34.9	25.6	44.8
R-C3D [34]	54.5	51.5	44.8	35.6	28.9	52.2
SSN [35]	*66.0*	*59.4*	51.9	41.0	29.8	-
A-Search [14]	-	-	51.8	42.4	30.8	-
CBR [36]	60.1	56.7	50.1	41.3	31.0	50.1
BSN + UNet [37]	-	-	53.5	45.0	36.9	-
Re-thinking F-RCNN [38]	59.8	57.1	53.2	48.5	42.5	-
D-SSAD [39]	-	-	*60.2*	*54.1*	*44.2*	*59.7*
Ours (TSN backbone)	-	-	-	-	-	62.4
Ours (I3D backbone)	-	-	-	-	-	**65.6**

- ActivityNet1.2

ActivityNet v1.2					
Approach	Detection mAP@				Spotting mAP
	0.5	0.75	0.95	Avg	
W-TALC [45]	37.0	14.6	-	18.0	-
SSN-SW [35]	-	-	-	18.1	-
3C-Net [46]	37.2	23.7	9.2	21.7	-
FPTADC [47]	37.6	21.8	2.4	21.9	-
SSN-TAG [35]	39.2	25.3	5.4	25.9	55.4
BSN [48]	46.5	30.0	8.0	30.0	49.6
BMN [49]	50.1	34.8	8.3	33.85	55.3
Ours (TSN backbone)	-	-	-	-	58.1
Ours (I3D backbone)	-	-	-	-	**60.2**

Conclusion

Key idea: **Use of Reinforcement Learning +**
End-to-End training

Key properties:

- Only need action detection ground-truth
- Able to sparsely browse videos
- mAP as training criterion

25th International Conference on Pattern Recognition

ActionSpotter: Deep Reinforcement Learning Framework for Temporal Action Spotting in Videos

Guillaume VAUDAUX-RUTH[1,2], Adrien CHAN-HON-TONG[1,3], Catherine[2] ACHARD

[1]ONERA [2]Sorbonne Université [3]Université Paris-Saclay

ONERA THE FRENCH AEROSPACE LAB

SORBONNE UNIVERSITÉ

université PARIS-SACLAY