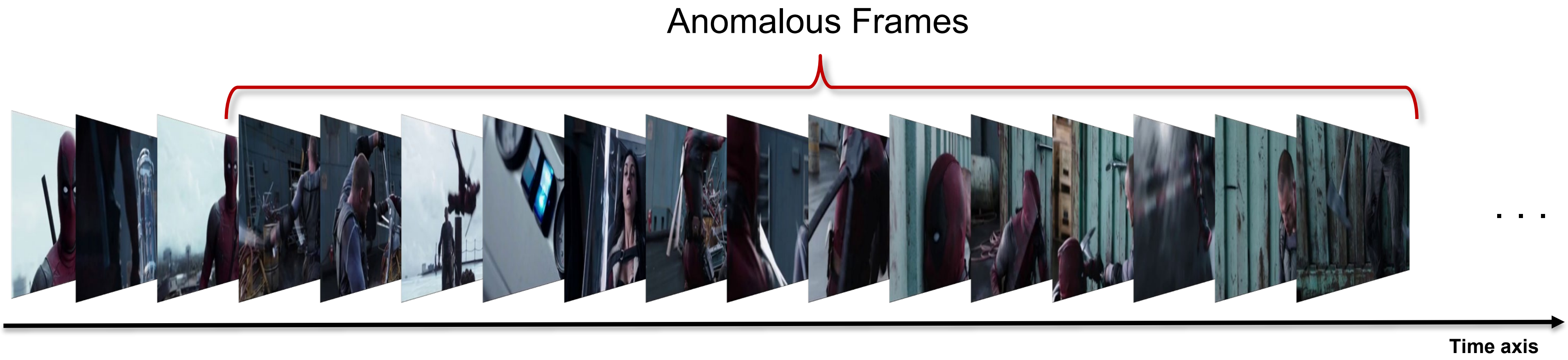
Consistency-based Self-supervised Learning for Temporal Anomaly Localization

European Conference on Computer Vision Workshops (ECCVW) – Tel Aviv, 2022

Aniello Panariello Angelo Porrello Simone Calderara Rita Cucchiara

AImageLab
University of Modena and Reggio Emilia, Italy

Video Anomaly Detection



- Video Anomaly Detection (VAD) is the task of detecting anomalous (human) activities in videos.
- Recent popular approaches are weakly supervised, in which videos are labeled at video level.

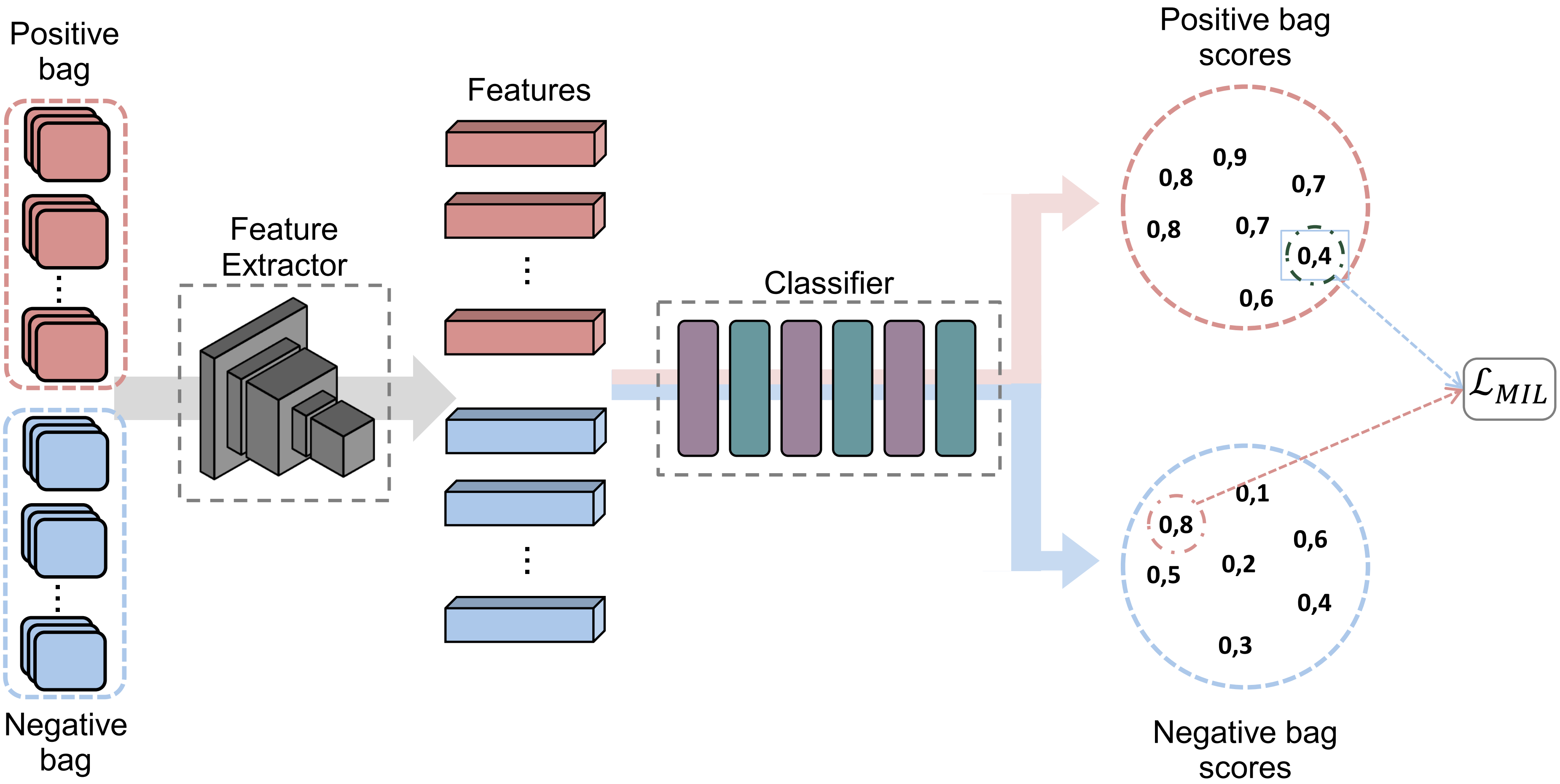
Standard video split

Videos get split in 32 frames clips.

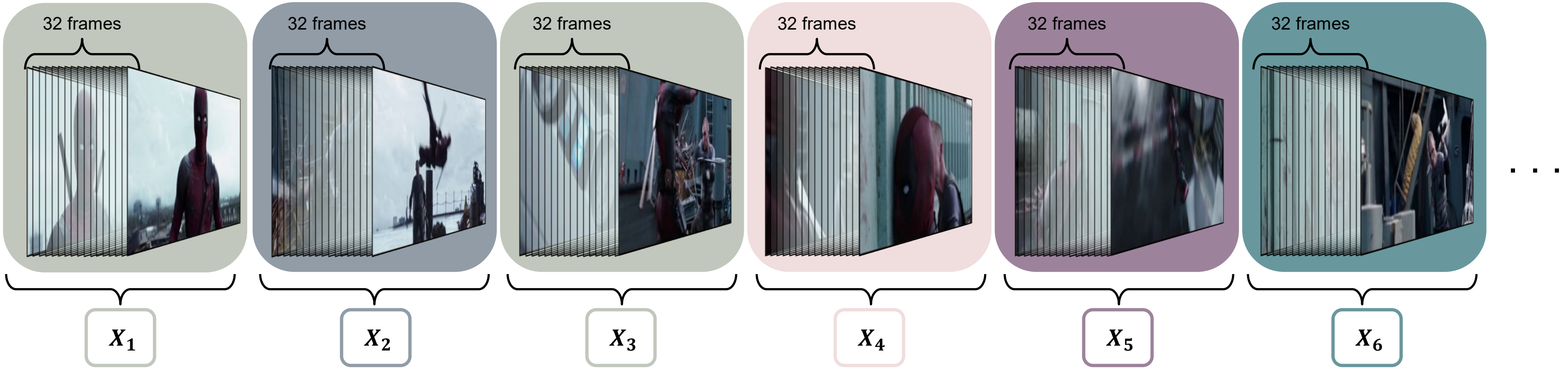
Standard video split: Every clip gets processed by the model.

In weakly-supervised methods, clips from anomalous videos are grouped into positive bags, while normal ones into negative bags.

MIL


Positive bag
Negative bag
Feature Extractor
Features
Classifier
Positive bag scores
0,9
0,8
0,7
0,8
0,7
0,4
0,6
$\mathcal{L}_{MIL}$
0,1
0,8
0,6
0,5
0,2
0,4
0,3
Negative bag scores

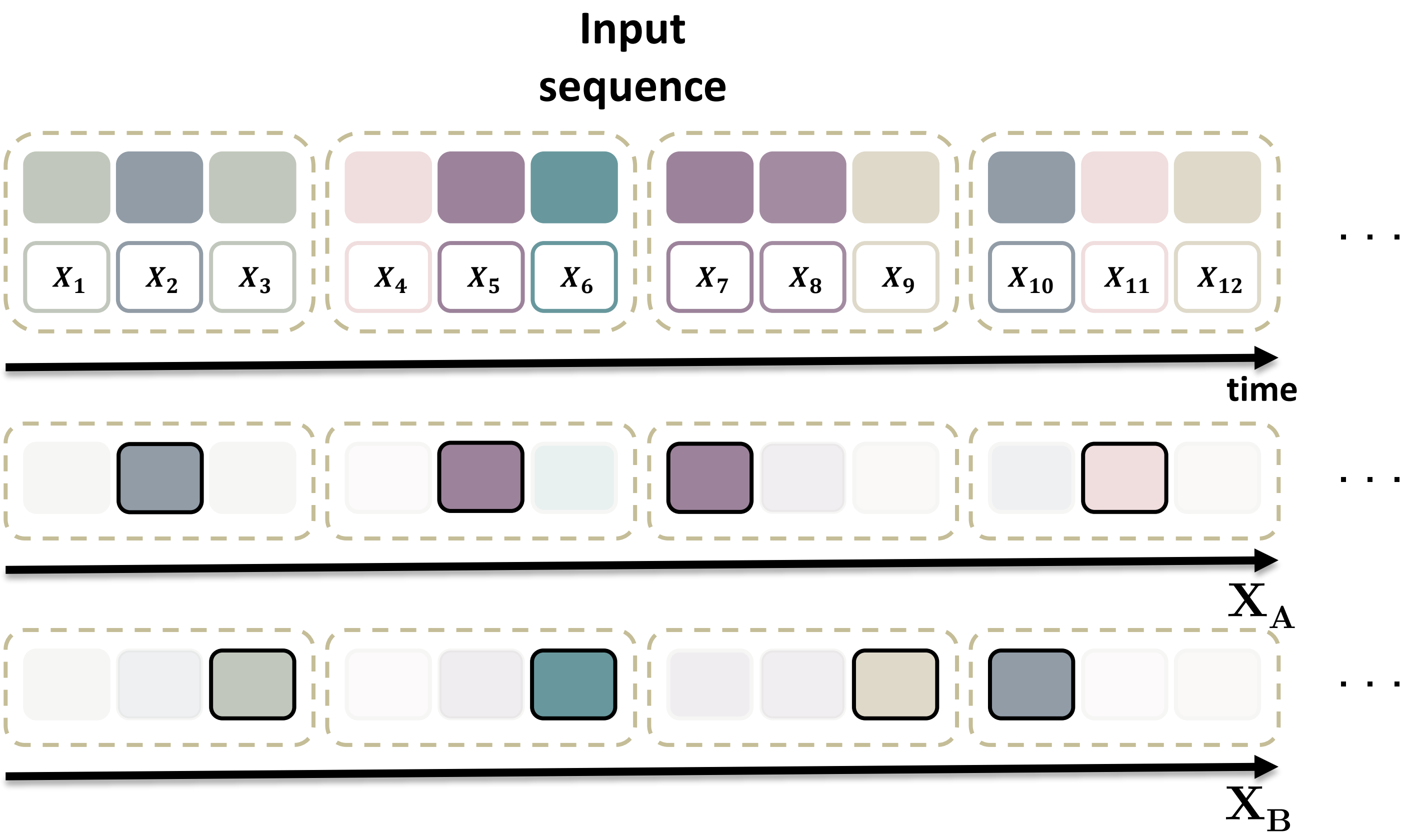
Custom Sampling



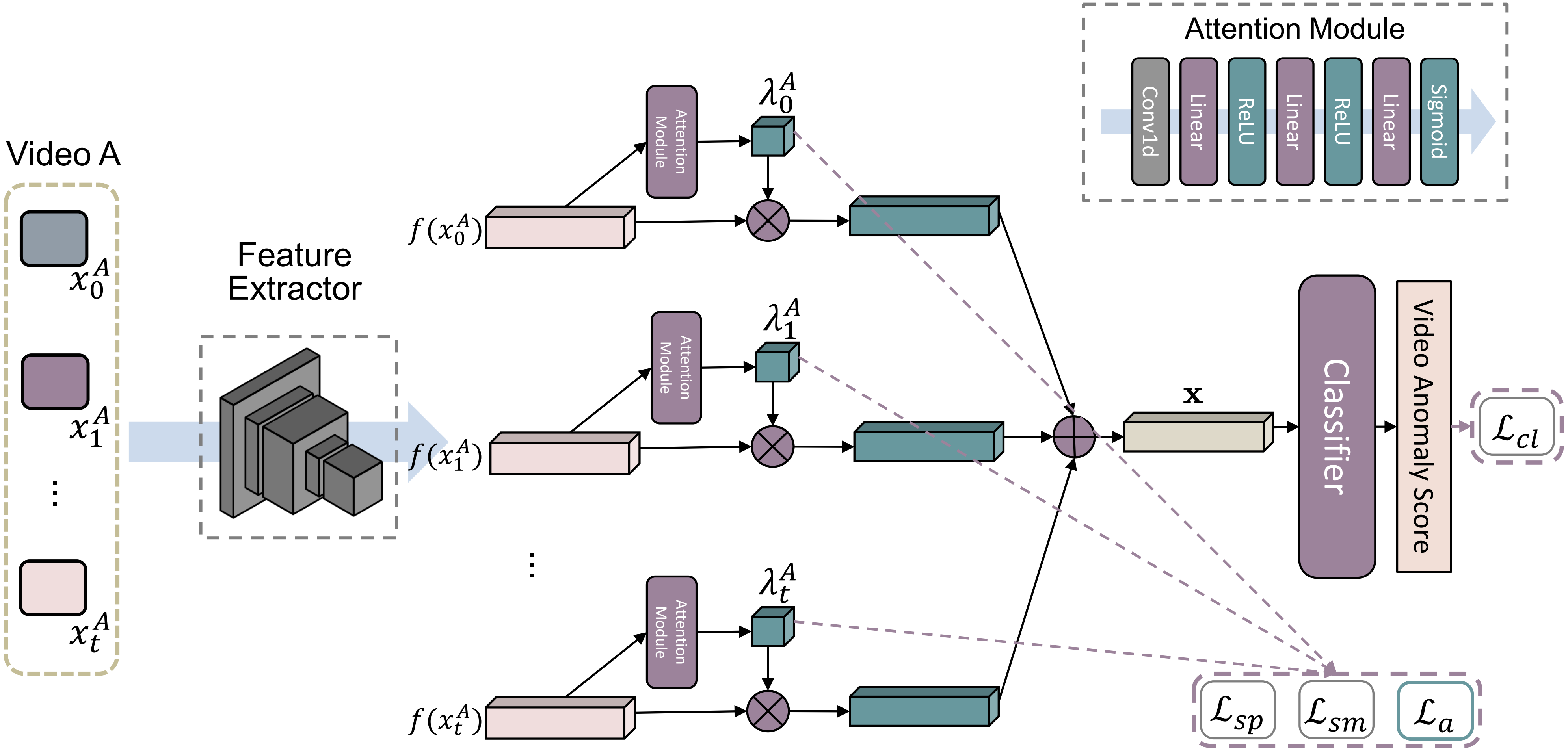
Videos get split in 32 frames clips.

Custom sampling: We sample only one clip from every window of three clips.

Custom Sampling



Model



Attention Module
Conv1d
Linear
ReLU
Linear
ReLU
Linear
Sigmoid
Video A
x_0^A
x_1^A
x_t^A
Feature Extractor
$f(x_0^A)$
$f(x_1^A)$
$f(x_t^A)$
Attention Module
λ_0^A
λ_1^A
λ_t^A
$\mathbf{x}$
Classifier
Video Anomaly Score
$\mathcal{L}_{cl}$
$\mathcal{L}_{sp}$
$\mathcal{L}_{sm}$
$\mathcal{L}_a$

Objective Function

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{sm} + \mathcal{L}_{sp} + \mathcal{L}_{a}$$

Objective Function

$$\mathcal{L} = \boxed{\mathcal{L}_{cl}} + \mathcal{L}_{sm} + \mathcal{L}_{sp} + \mathcal{L}_{a}$$

Classifier output

$$\mathcal{L}_{cl} = \mathrm{BCE}(\boxed{f(\mathcal{X}_i)}, \boxed{y_i})$$

True video label

The **classification loss** is a binary cross-entropy at video-level.

This is used to ensure the model can recognize if a video contains anomalies or not.

Objective Function

$$\mathcal{L} = \mathcal{L}_{cl} + \boxed{\mathcal{L}_{sm}} + \mathcal{L}_{sp} + \mathcal{L}_{a}$$

$$\mathcal{L}_{cl} = \mathrm{BCE}(f(\mathcal{X}_i), y_i)$$

$$\mathcal{L}_{sm} = \sum_{t=1}^{T-1} (\lambda_t - \lambda_{t+1})^2$$

The **smooth loss** impose adjacent attention coefficients to vary as little as possible.

Objective Function

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{sm} + \boxed{\mathcal{L}_{sp}} + \mathcal{L}_a$$

$$\mathcal{L}_{cl} = \mathrm{BCE}(f(\mathcal{X}_i), y_i)$$

$$\mathcal{L}_{sm} = \sum_{t=1}^{T-1}(\lambda_t - \lambda_{t+1})^2$$

$$\mathcal{L}_{sp} = || \lambda_t ||_1$$

The **sparsity loss** penalizes the ℓ_1 norm of the attention weights.
This reflects the rarity of the anomalies in videos.

Objective Function

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{sm} + \mathcal{L}_{sp} + \boxed{\mathcal{L}_{a}}$$

$$\mathcal{L}_{cl} = \mathrm{BCE}(f(\mathcal{X}_i), y_i)$$

$$\mathcal{L}_{sm} = \sum_{t=1}^{T-1}(\lambda_t - \lambda_{t+1})^2$$

$$\mathcal{L}_{sp} = ||\ \lambda_t\ ||_1$$

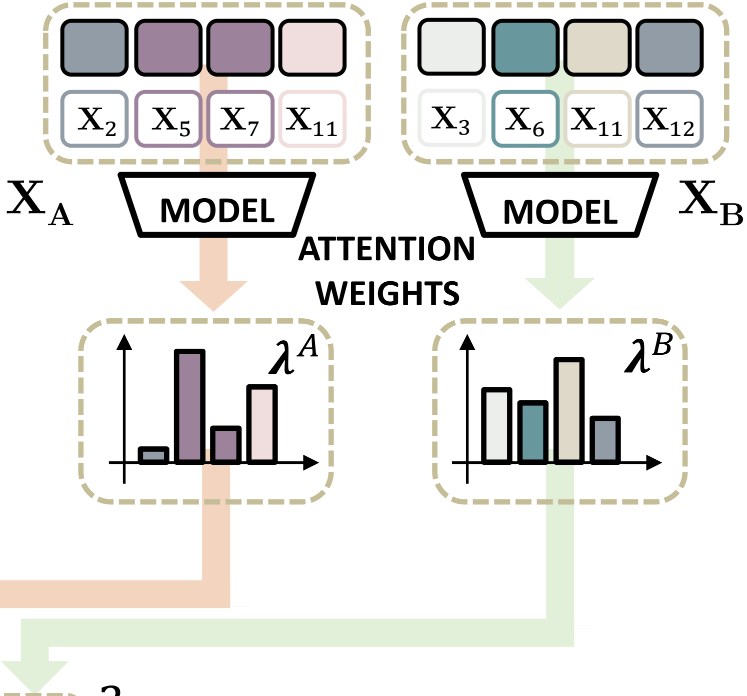
$$\mathcal{L}_{a} = \sum_{t=1}^{T}\left(\lambda_t^A - \lambda_t^B\right)^2$$

The alignment loss is a consistency-based regularization term.

Alignment Loss

1. Sample two slightly different version of the video.
2. Classify them and extract the attention weights of the clips.
3. The attention weights of the two sampling should be close together.

With this additional regularization term we enforce smoothness over a wider temporal horizon.



$$\mathcal{L}_a = \sum_{t=1}^{T} \left(\boxed{\lambda_t^A} - \boxed{\lambda_t^B}\right)^2$$

Quantitative Results

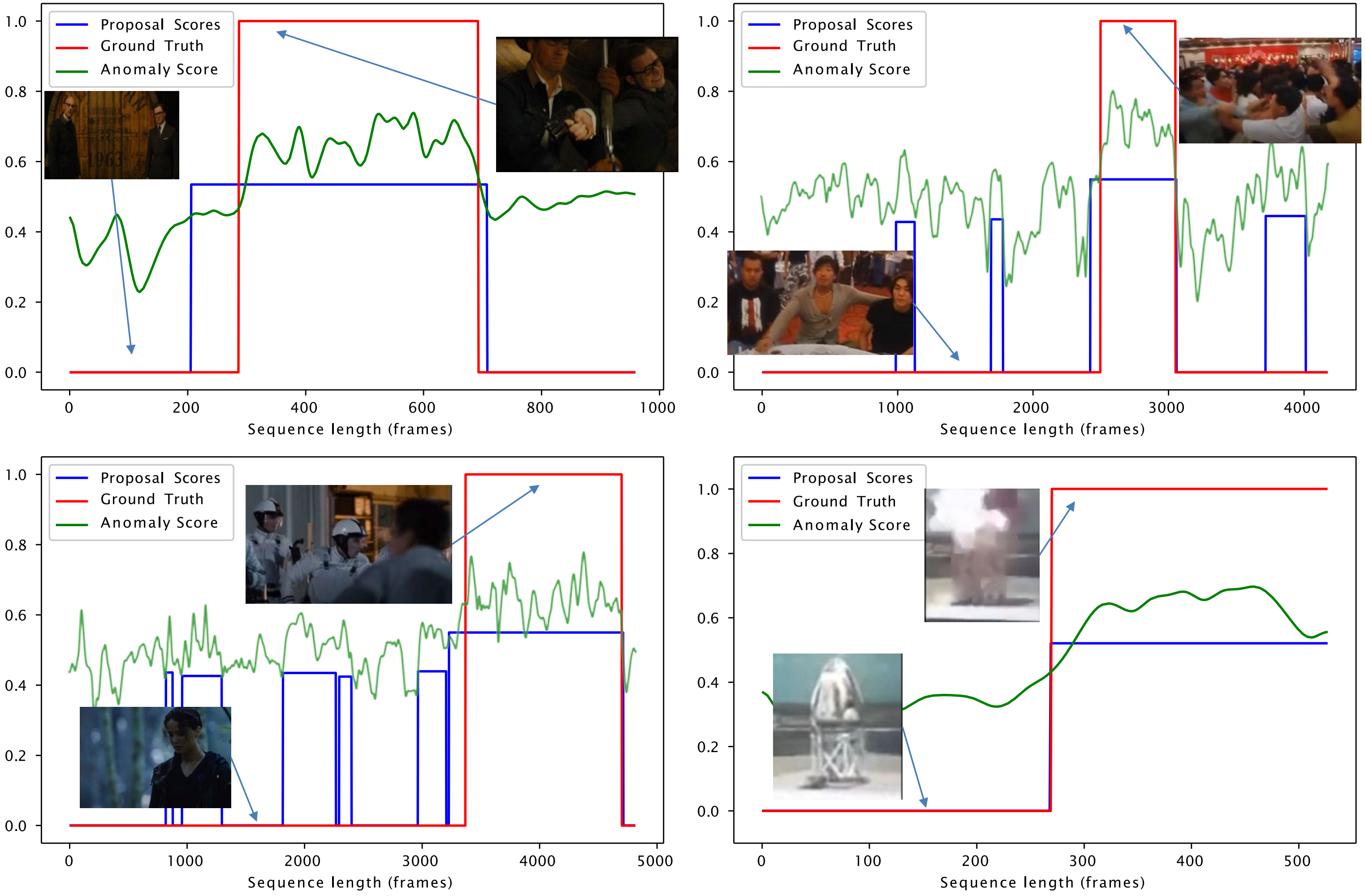
ECCV TEL AVIV 2022 AImage Lab University of Modena and Reggio Emilia

	Video Level		Segment Level		Frame Level Proposal		Frame Level	
Align Loss	AUC%	AP%	AUC%	AP%	AUC%	AP%	AUC%	AP%
-	**97.91**	**98.36**	84.39	66.75	85.14	68.01	84.57	65.96
✓	97.79	98.28	**85.49**	**66.87**	**90.23**	**71.68**	**85.65**	**66.05**

Qualitative Results

ECCV TEL AVIV 2022

AImage Lab University of Modena and Reggio Emilia



Conclusions

- Alignment loss allows to learn effective frame-level scores in weakly supervised settings.
- A base network equipped also with other common regularization techniques brings even more imporvements.

Code: [github.com]