MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection



Jia-Chang Feng *Fa-Ting Hong* *Wei-Shi Zheng*

School of Computer Science and Engineering, Sun Yat-sen University

Peng Cheng Laboratory

Key Laboratory of Machine Intelligence and Advanced Computing, Mistry of Education of China

Pazhou Lab

Speaker: Jia-Chang Feng

Video Anomaly Detection (VAD)

Project Page

- What is anomaly in the video?
  - Events that betray pre-defined patterns.
  - Events that catch the user’s interest, especially those related to crime.

Abnormal object

Strange action

Fighting

Road Accident

Figures are samples from CUHK-Avenue (left), UCF-Crime (right).

Video Anomaly Detection (VAD)

Project Page

- Real-world applications of VAD
  - Intelligent surveillance systems
  - Traffic monitoring / Smart City
  - Industrial monitoring systems
  - …

https://www.techtoreview.com/upload/1583312149.jpeg

https://www.aindralabs.com/wp-content/uploads/2.png

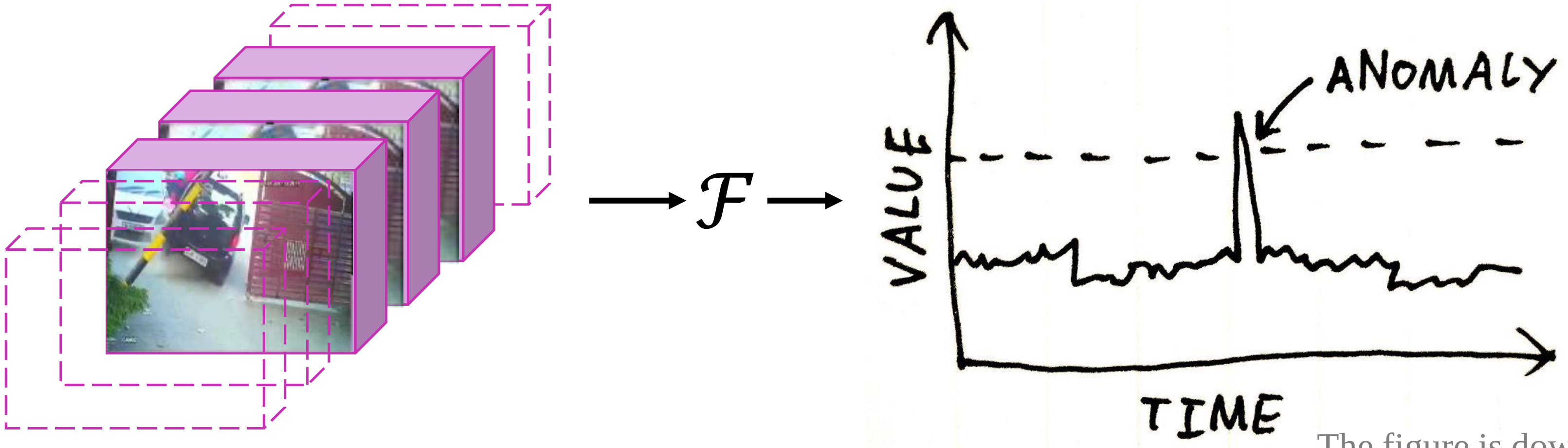
Project Page

Weakly Supervised VAD

- Heavy annotation cost for fully supervised VAD
- Scene bias / Inflexibility for unsupervised VAD

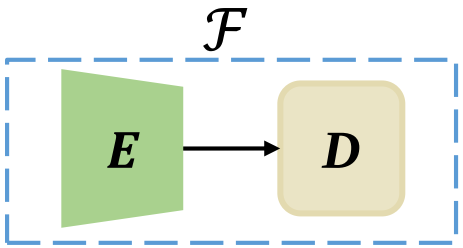
Weakly Supervised VAD

- Formulation:
  - Given a binary video-level label for each training video [Abnormal / Normal]
  - Train a prediction function $\mathcal{F}$ to predict frame-level anomaly scores



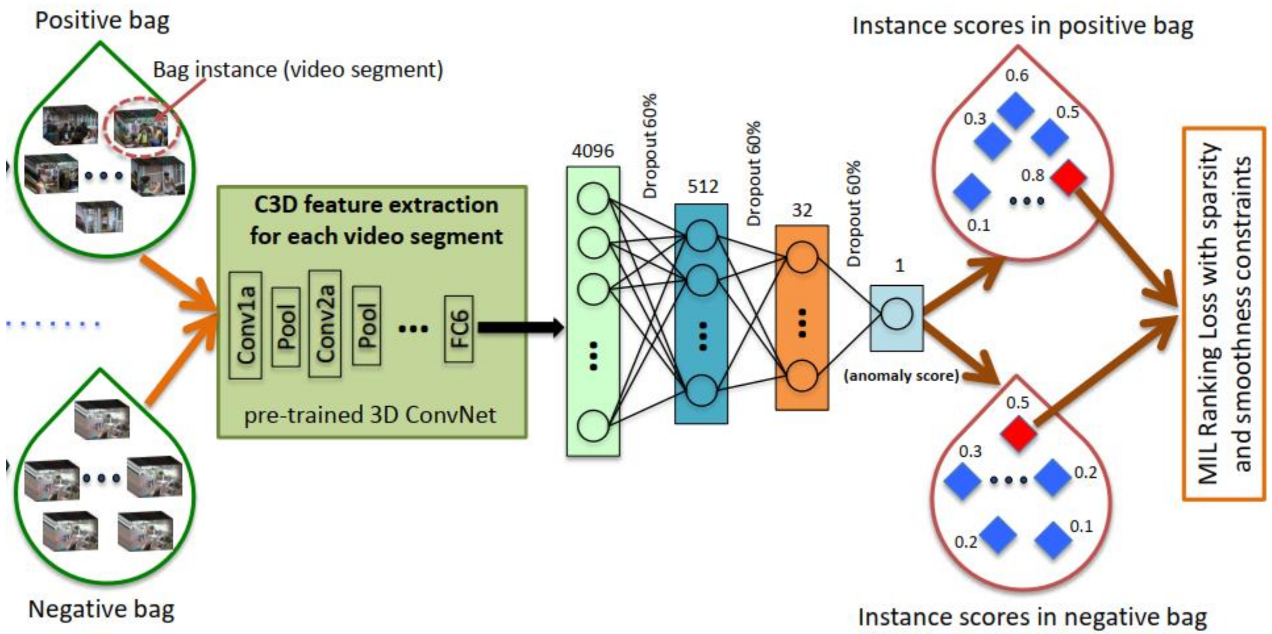
The figure is downloaded from the Internet.

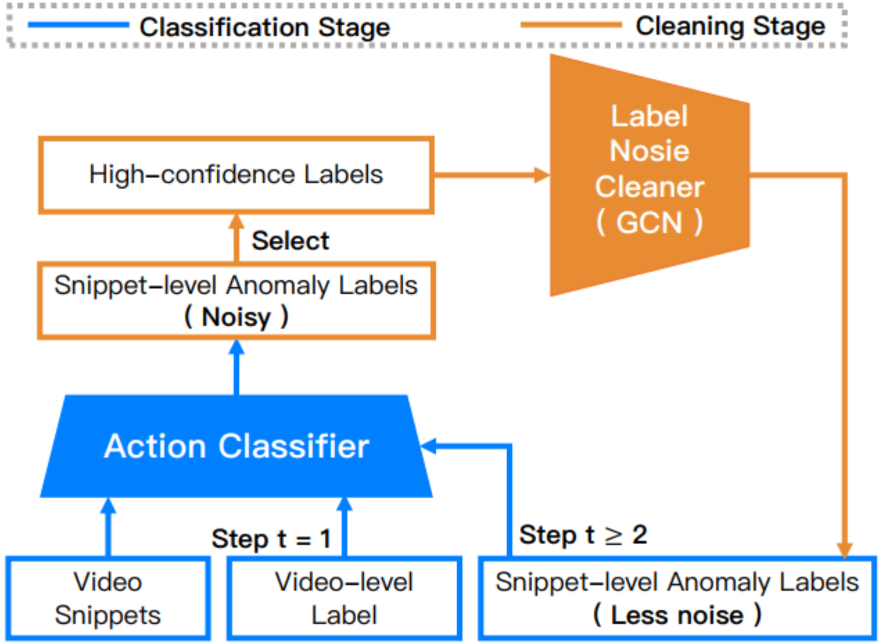
Weakly Supervised VAD



Project Page

- Previous works
  - Encoder-agnostic: **Train** specific-designed detector **only**
  - Encoder-based: **Train both** feature encoder and specific-designed detector



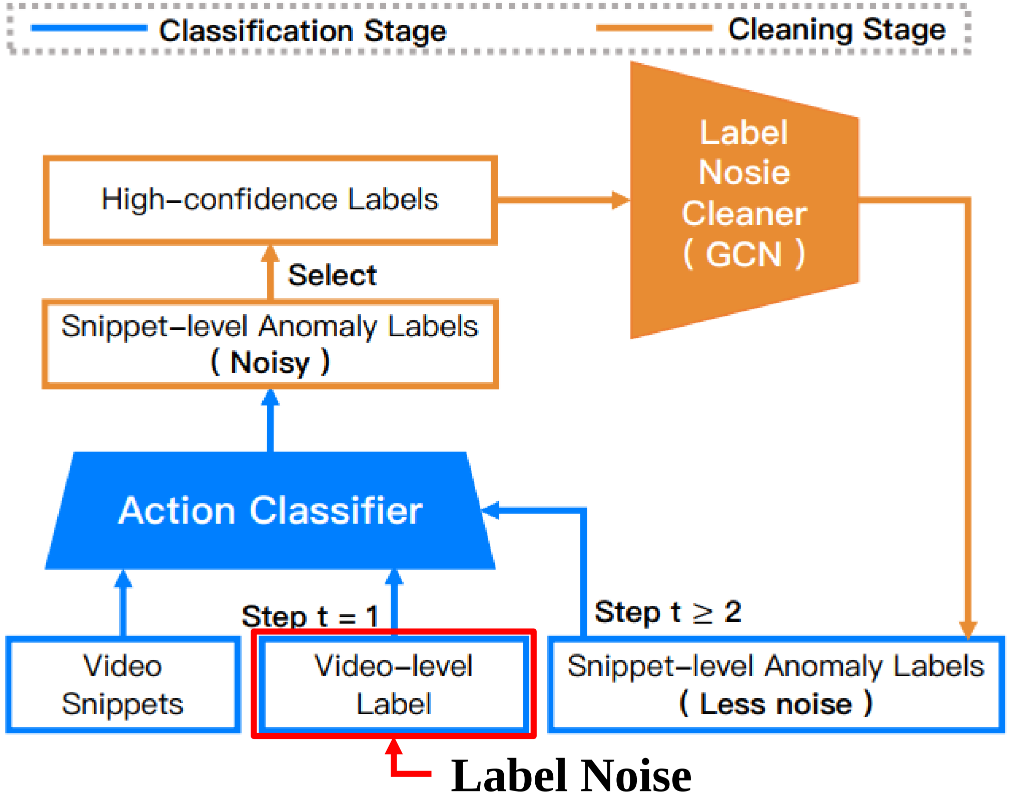


The figure (left) is from Sultani .et al (CVPR 2018), while the right is from Zhong .et al. (CVPR 2019).

Weakly Supervised VAD

Project Page

- Problems in previous encoder-based method [Zhong .et al, 2019]
  - Label noise is introduced in the first iteration.
  - High training computation cost
    - Multiple training iterations
  - High testing computation cost
    - 10-crop testing augmentation



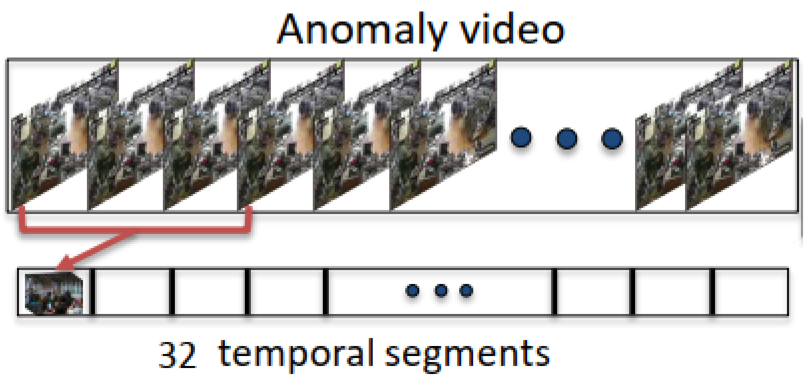
The figure is from Zhong .et al (CVPR 2019)

Multiple Instance Self-Training Framework

Project Page

- Motivations
  - To minimize the domain gap lying between common videos and surveillance videos.
    - Small foreground
    - Not actor-centric
    - Blur
  - Online fine-grained anomaly detection
  - Spatial anomaly explanation/localization

b) Samples of UCF-Crime

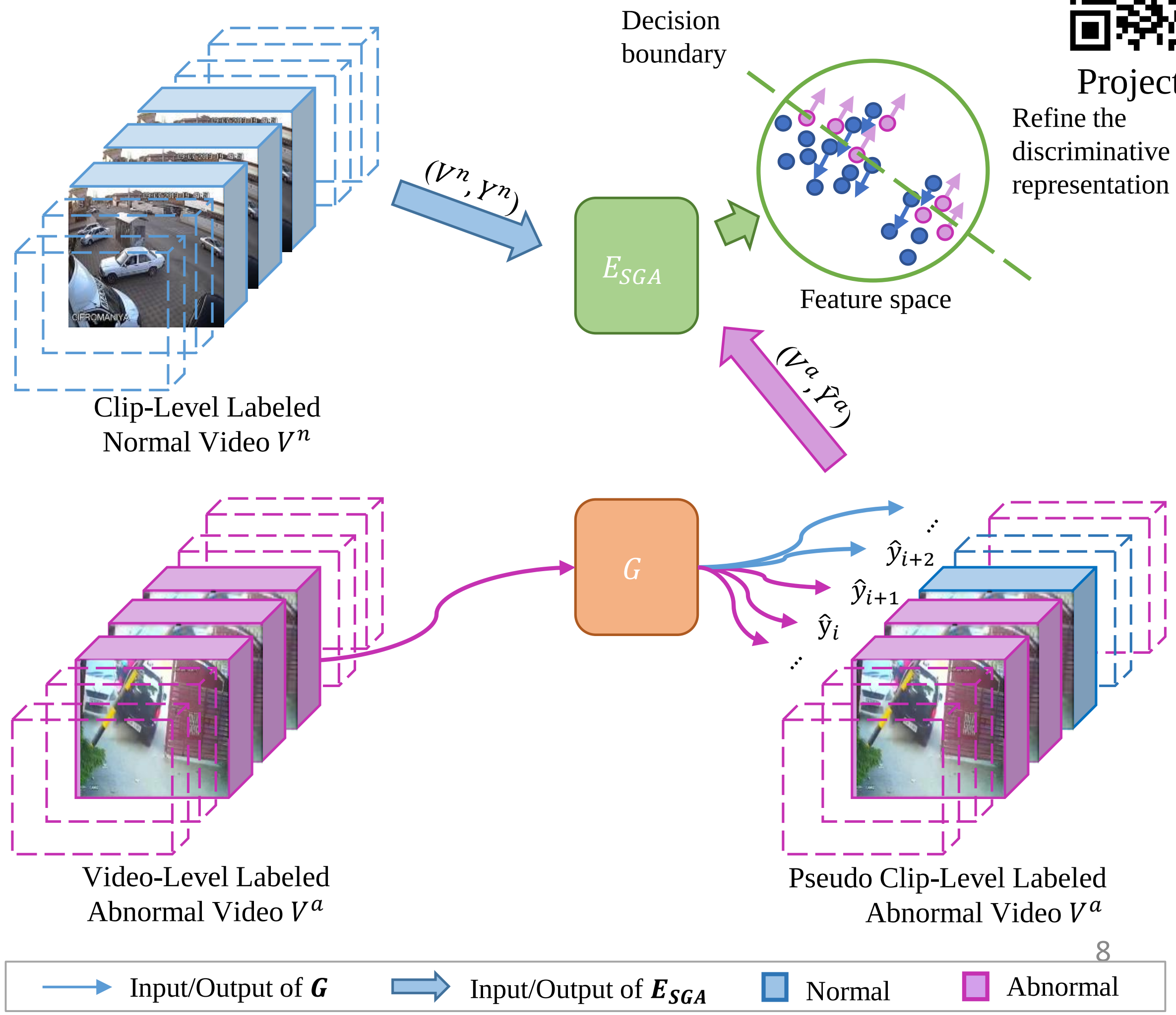


The figure is from Sultani .et al (CVPR 2018)

Multiple Instance Self-Training Framework

- Contributions:
  - **An online fine-grained method** for weakly supervised VAD.
  - **An efficient way** to finetune feature encoder with a two-stage framework.
  - **Sparse-continuous sampling strategy** for MIL-based pseudo label generator.
  - **Self-Guided Attention Module** to enhance the feature encoder.
  - MIST provide **spatial explanation / visualization** for anomalous events.

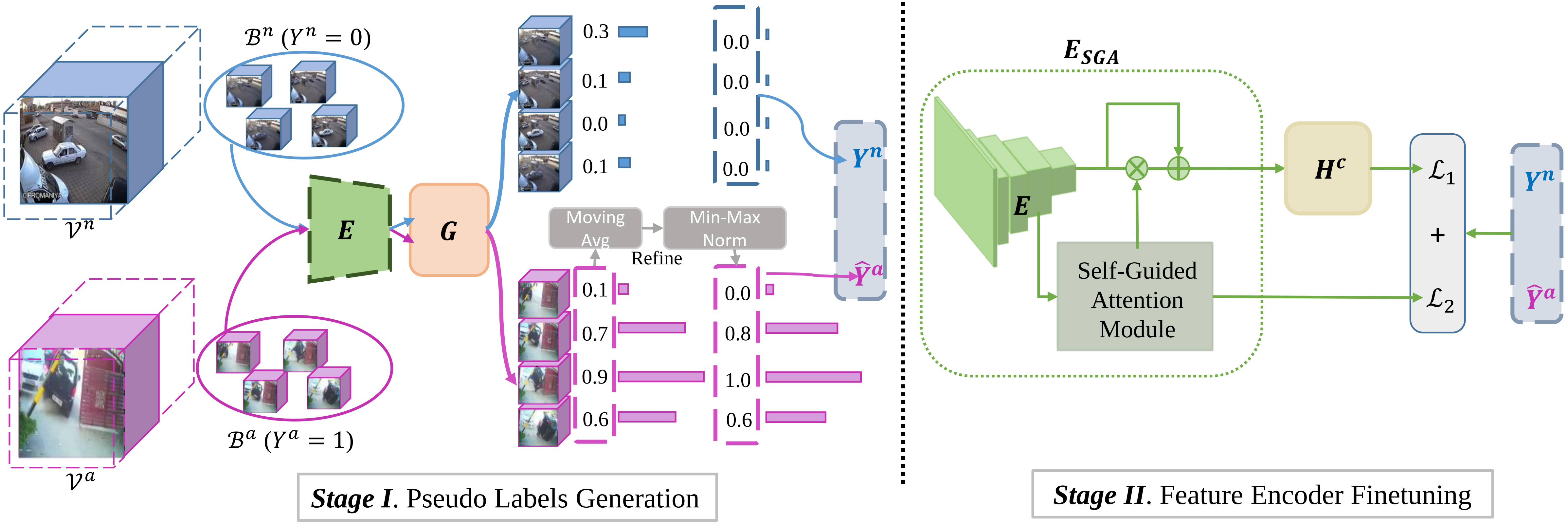
Project Page



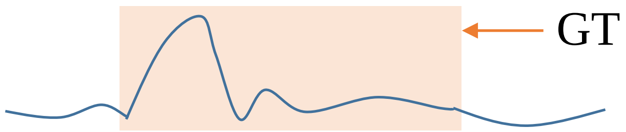
Multiple Instance Self-Training Framework

Project Page

- Overview



Stage I: Pseudo Labels Generation

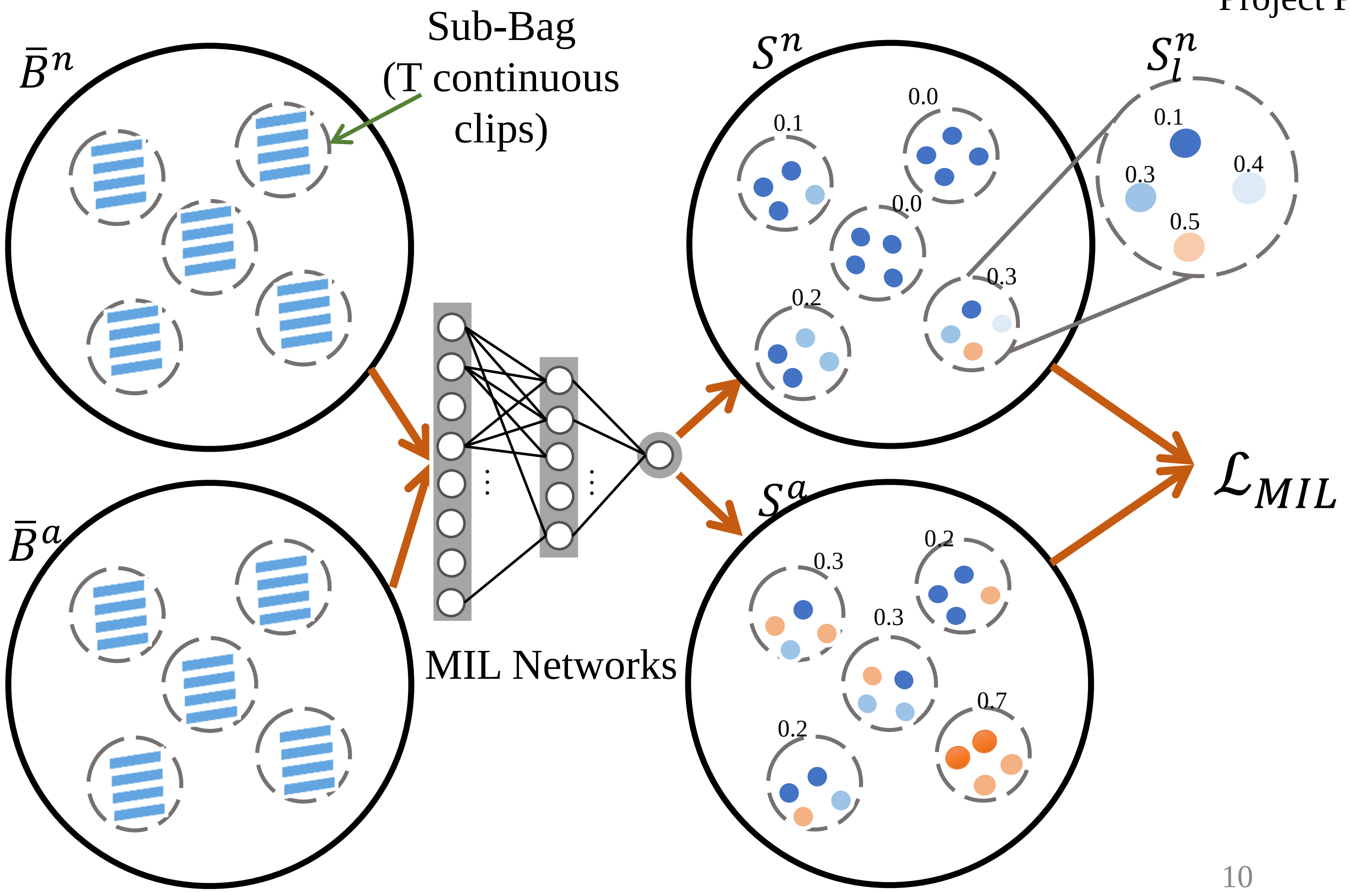


Project Page

- Sparse Continuous Sampling:
  - Uniformly sampling L sub-bag
    - Sub-bag: T continuous clips

- Training Objective:

$$\mathcal{L}_{MIL} = \left(\epsilon - \max_{1\le l\le L} \mathcal{S}_l^a + \max_{1\le l\le L} \mathcal{S}_l^n\right)_+ + \frac{\lambda}{L}\sum_{l=1}^{L}\mathcal{S}_l^a.$$



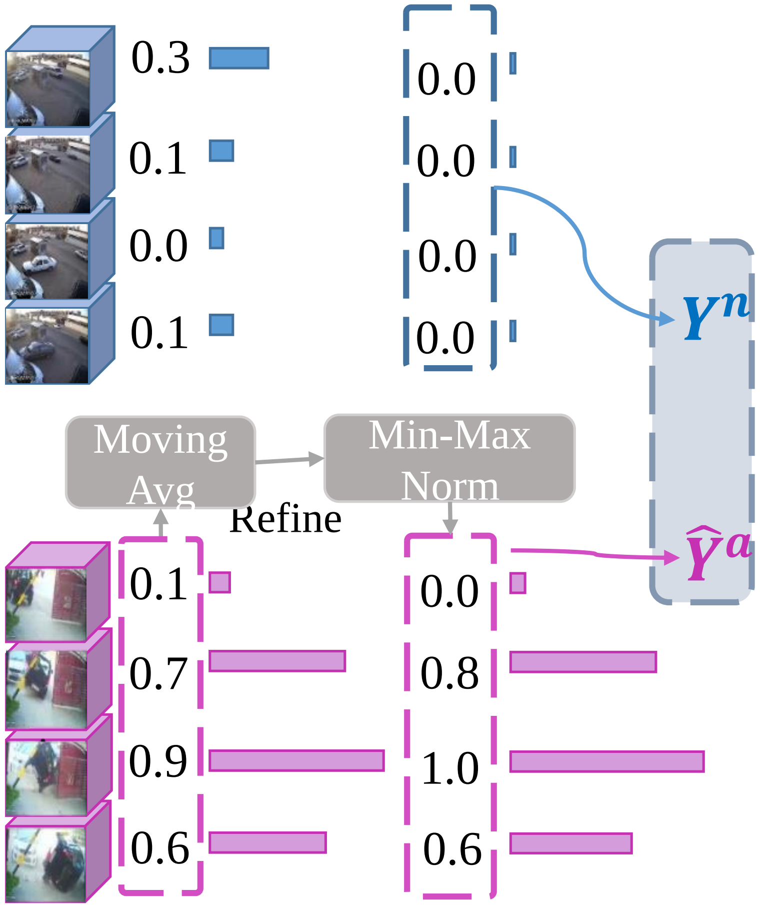
Stage I: Pseudo Labels Generation

- Pseudo Labels Refinement
  - Moving Average Smoothing

$$\tilde{s}_i^a = \frac{1}{2k} \sum_{j=i-k}^{i+k} s_j^a$$

  - Min-Max Normalization

$$\hat{y}_i^a = \left(\tilde{s}_i^a - \min \tilde{S}^a\right) / (\max \tilde{S}^a - \min \tilde{S}^a)), i \in [1, N]$$

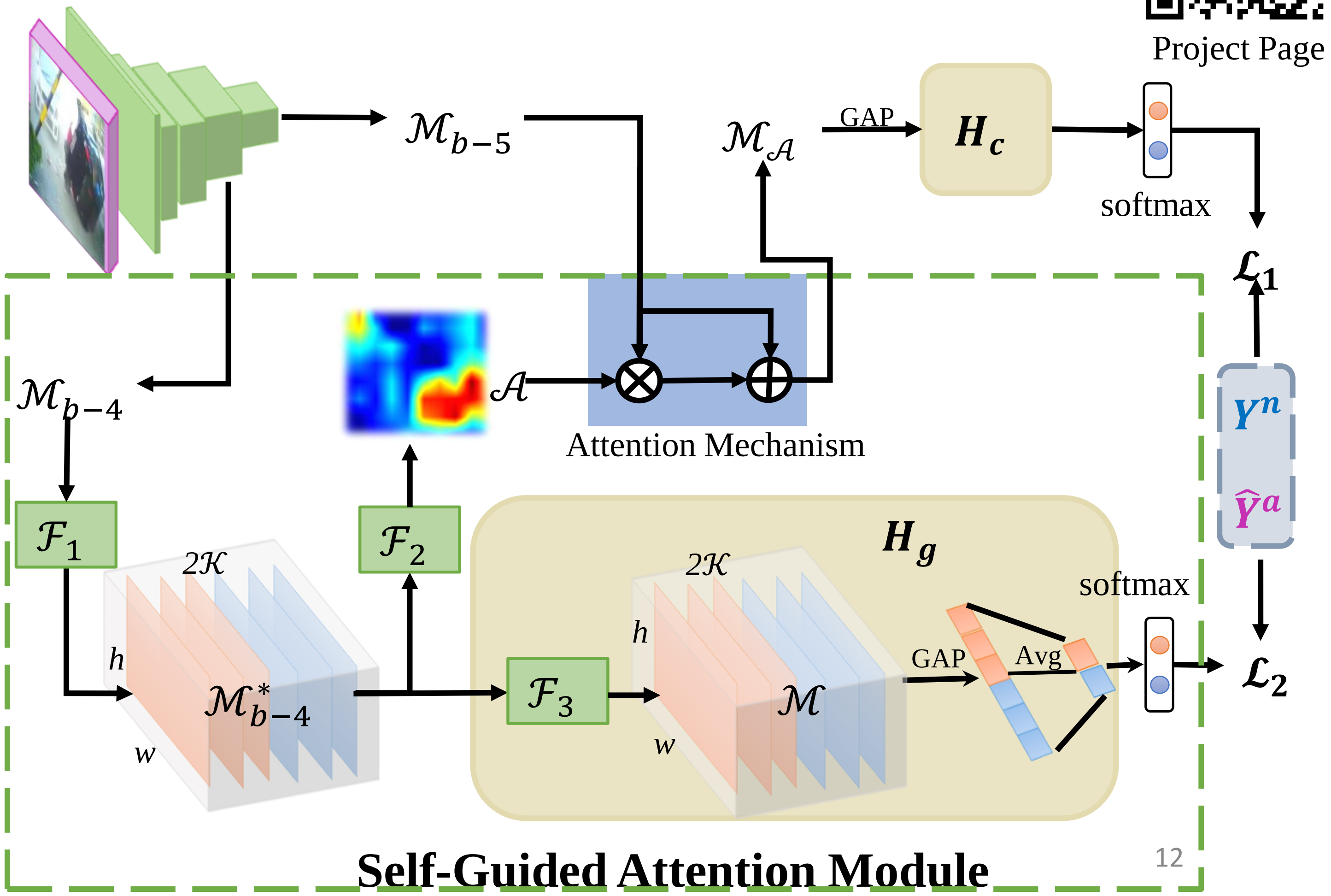


Project Page

Stage II: Feature Encoder Finetuning

- Self-Guided Attention Module:
  - Attention generation
    $\mathcal{A} = \mathcal{F}_2(\mathcal{F}_1(\mathcal{M}_{b-4}))$
  - Attention Mechanism
    $\mathcal{M}_{\mathcal{A}} = \mathcal{M}_{b-5} + \mathcal{A} \circ \mathcal{M}_{b-5}$
  - Indirect guidance by a guided classification head $\boldsymbol{H}_g$ to make $\mathcal{M}^*_{b-4}$ more discriminative.

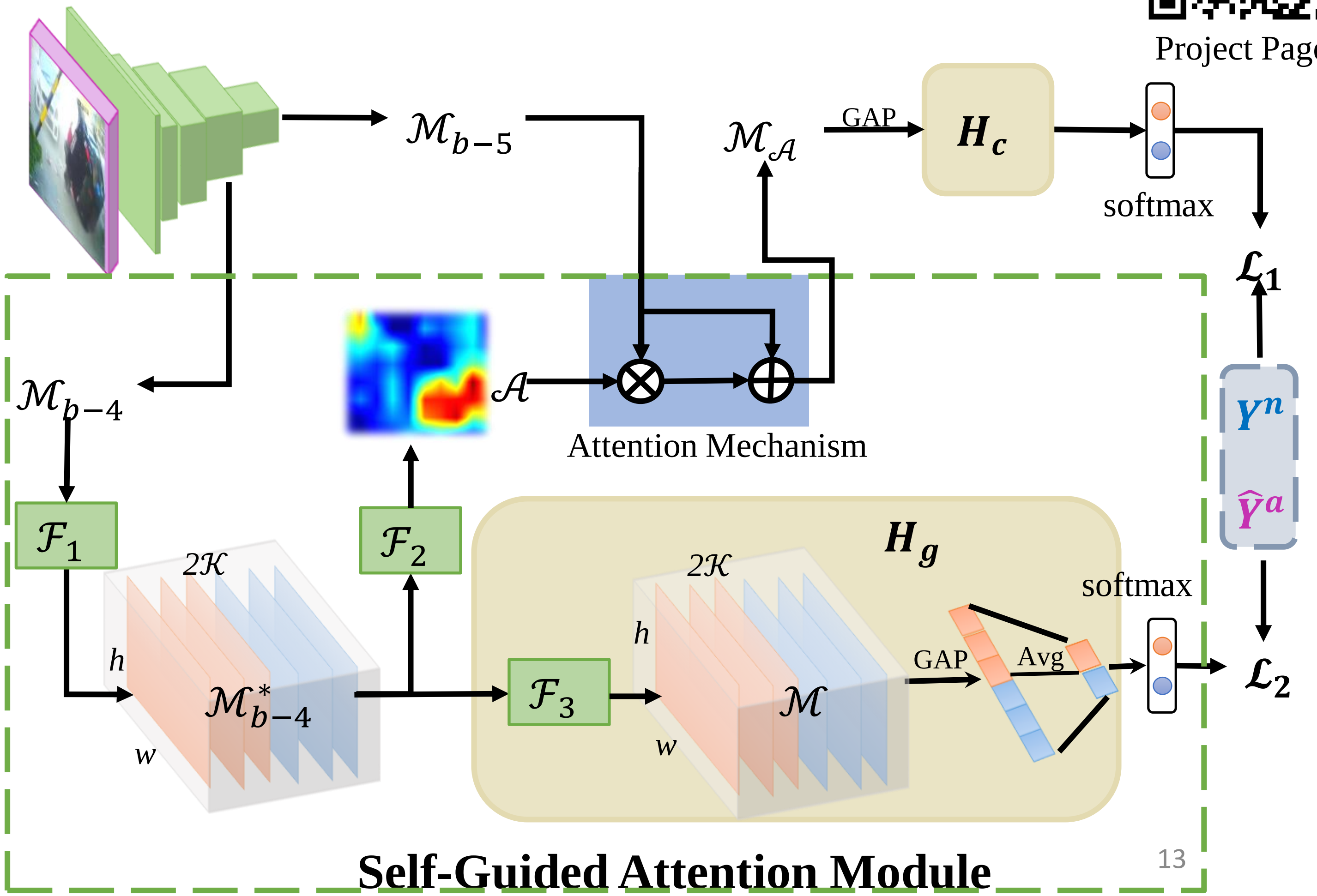
Project Page



Stage II: Feature Encoder Finetuning

- Training Objective in Finetuning
  - $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$
  - $\mathcal{L}_1, \mathcal{L}_2$: class-weighted cross-entropy loss $\mathcal{L}_w$

$$\mathcal{L}_w = -w_0 y \log p - w_1 (1-y) \log(1-p),$$

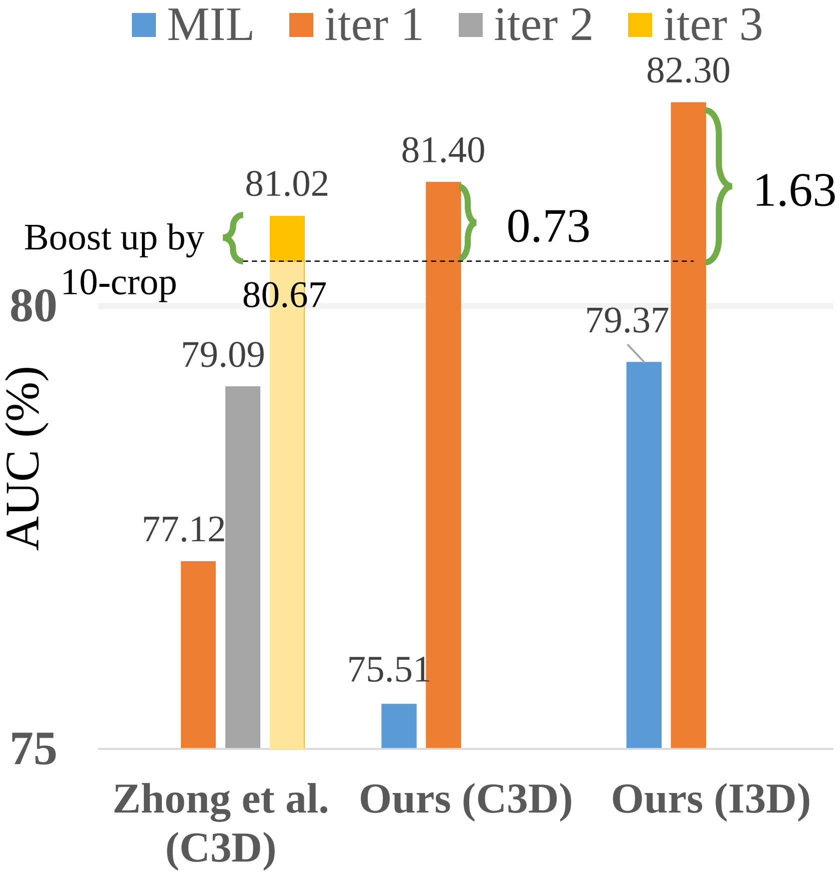


Experimental Results

Method	Supervised	Grained	Encoder	AUC (%)	FAR (%)
Hasan et al. [7]	Un	Coarse	AE^{RGB}	50.6	27.2
Lu et al. [16]	Un	Coarse	Dictionary	65.51	3.1
SVM	Weak	Coarse	$C3D^{RGB}$	50	-
Sultani et al. [23]	Weak	Coarse	$C3D^{RGB}$	75.4	1.9
Zhang et al. [32]	Weak	Coarse	$C3D^{RGB}$	78.7	-
Zhu et al. [38]	Weak	Coarse	AE^{Flow}	79.0	-
Zhong et al. [35]	Weak	Fine	$C3D^{RGB}$	80.67*(81.08)	3.3*(2.2)
Liu et al. [13]	Full(T)	Fine	$C3D^{RGB}$	70.1	-
Liu et al. [13]	Full(S+T)	Fine	NLN^{RGB}	82.0	-
MIST	Weak	Fine	$C3D^{RGB}$	**81.40**	2.19
MIST	Weak	Fine	$I3D^{RGB}$	**82.30**	**0.13**

Table 1: Quantitative comparisons with existing online methods on UCF-Crime under different levels of supervision and fineness of prediction. The results in $(\cdot)$ are tested with *10-crop*, while those marked by * are tested without.

Project Page

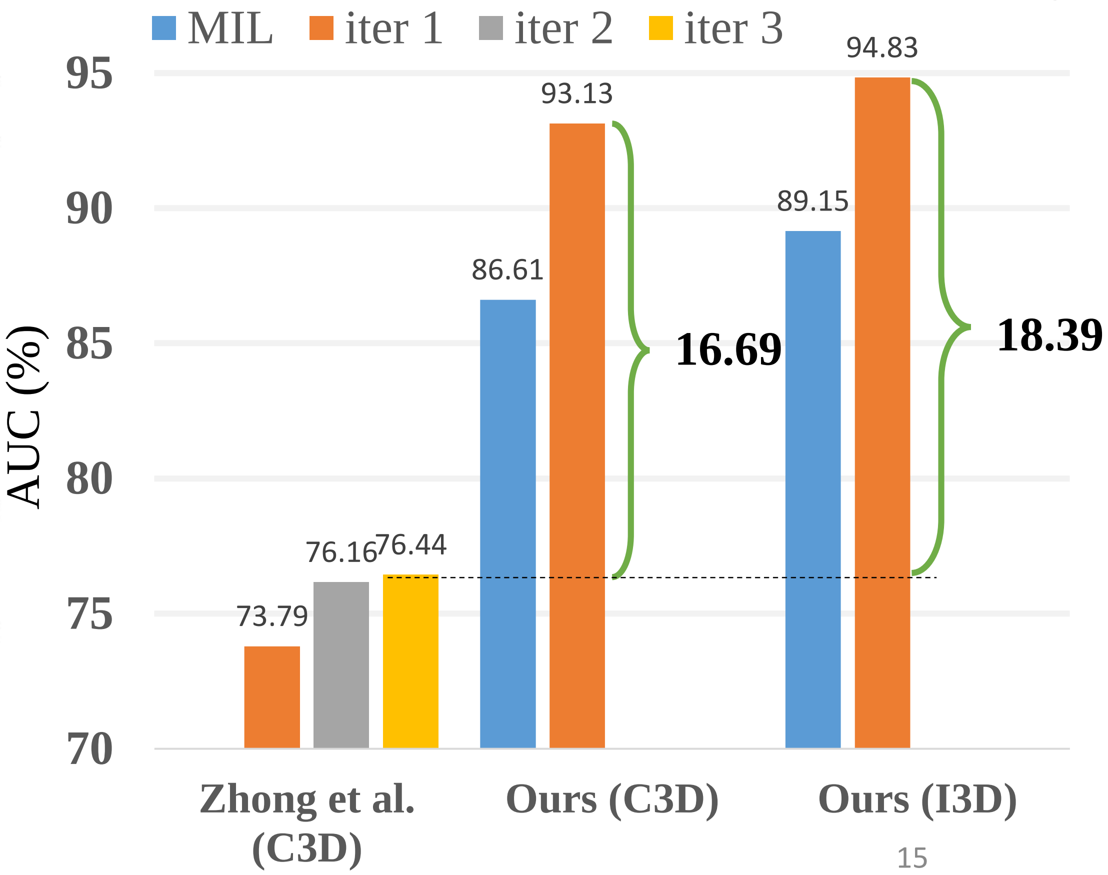


Experimental Results

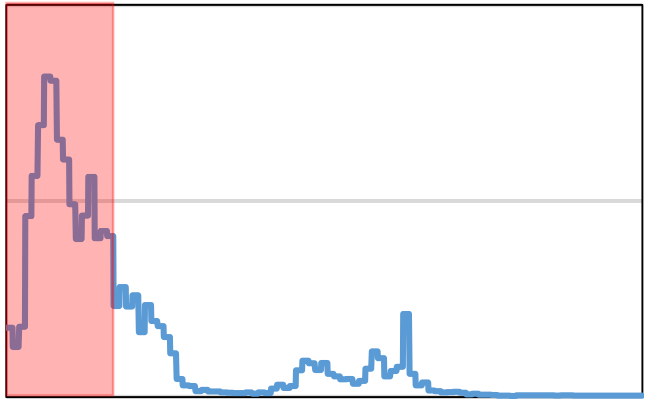
Project Page

Method	Feature Encoder	Grained	AUC (%)	FAR (%)
Sultani *et al.* [23]	$C3D^{RGB}$	Coarse	86.30	0.15
Zhang *et al.* [32]	$C3D^{RGB}$	Coarse	82.50	0.10
Zhong *et al.* [35]	$C3D^{RGB}$	Fine	76.44	-
AR-Net [27]	$C3D^{RGB}$	Fine	85.01*	0.57*
AR-Net [27]	$I3D^{RGB}$	Fine	85.38	0.27
AR-Net [27]	$I3D^{RGB+Flow}$	Fine	91.24	0.10
MIST	$C3D^{RGB}$	Fine	**93.13**	1.71
MIST	$I3D^{RGB}$	Fine	**94.83**	**0.05**

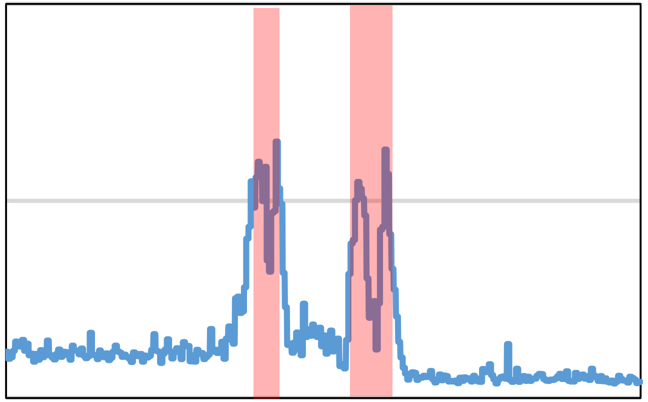
Table 2: Quantitative comparisons with existing methods on ShanghaiTech. The results with * are re-implemented.



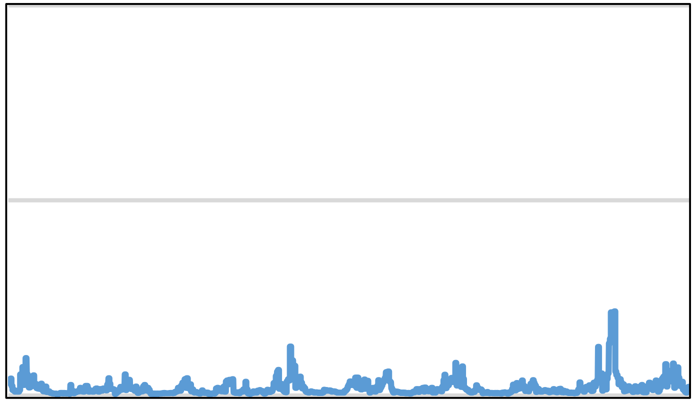
Temporal Visualization on UCF-Crime



Shooting008

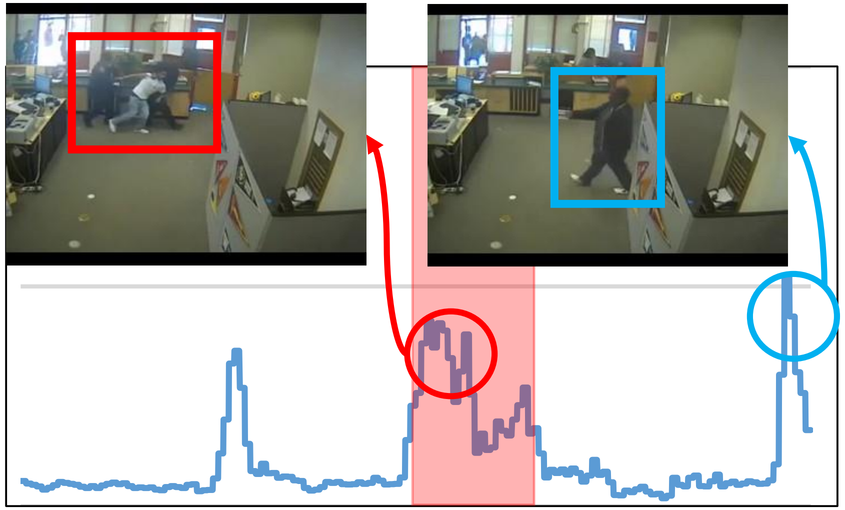


Vandalism028

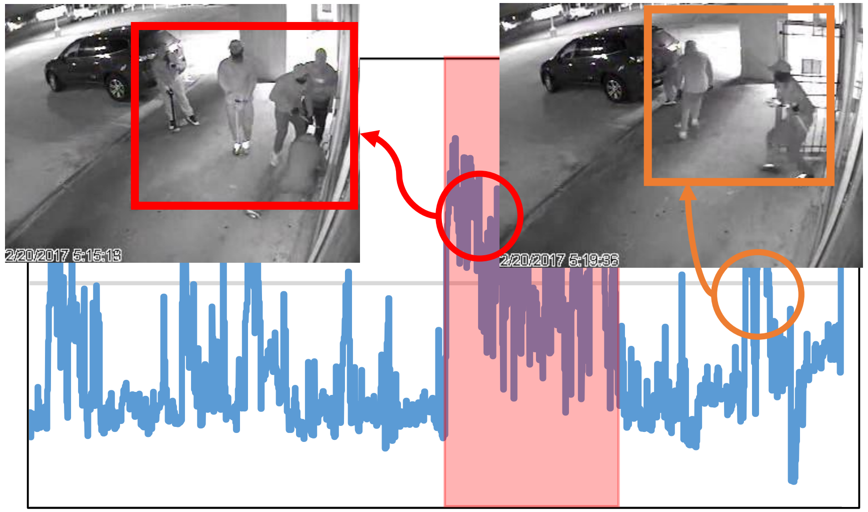


Normal877

Project Page



Arrest001



Burglary079

Effect of MIST finetuning

Project Page

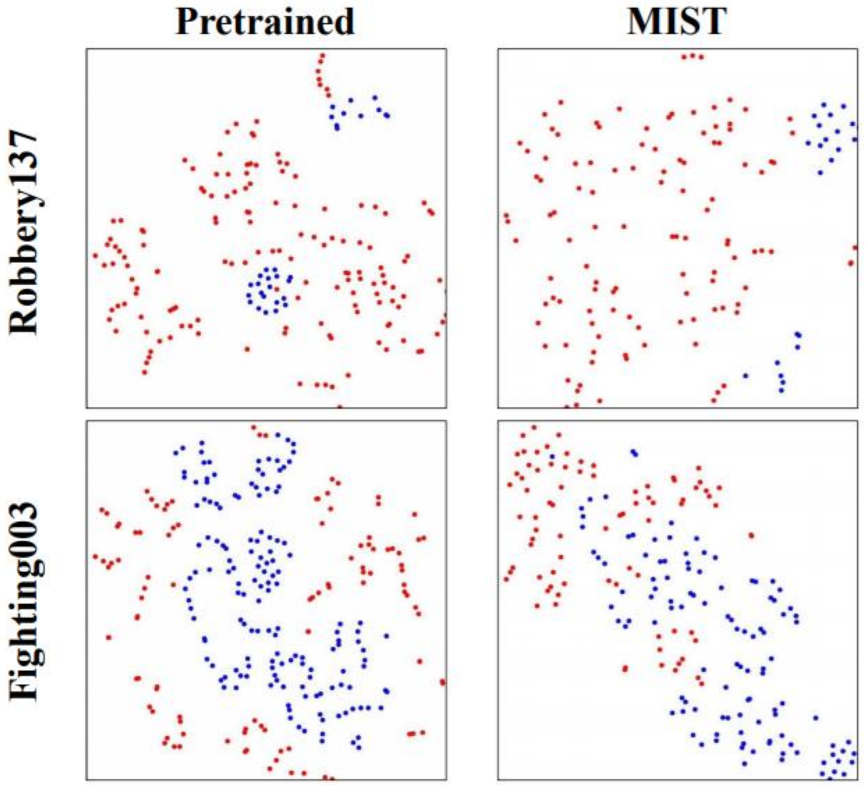


Figure 6: Feature space visualization of pretrained vanilla feature encoder **I3D** and the MIST fine-tuned encoder via t-SNE [18] on UCF-Crime testing videos. The red dots denote anomalous regions while the blue ones are normal.

Encoder-Agnostic Methods	AUC (%)					
	UCF-Crime			ShanghaiTech		
	pretrained	MIST	Δ	pretrained	MIST	Δ
Sultani *et al.* [20]	78.43	81.42	**+2.99**	86.92	92.63	**+5.71**
Zhang *et al.* [28]	78.11	81.58	**+3.47**	88.87	92.50	**+3.63**
AR-Net [24]	78.96	82.62	**+3.66**	85.38	92.27	**+6.89**
Our MIL generator	79.37	81.55	**+2.18**	89.15	92.24	**+3.09**

Table 3: Quantitative comparisons between the features from the pretrained vanilla feature encoder and those from MIST on UCF-Crime and ShanghaiTech datasets by adopting encoder-agnostic methods.

Ablation Studies

Project Page

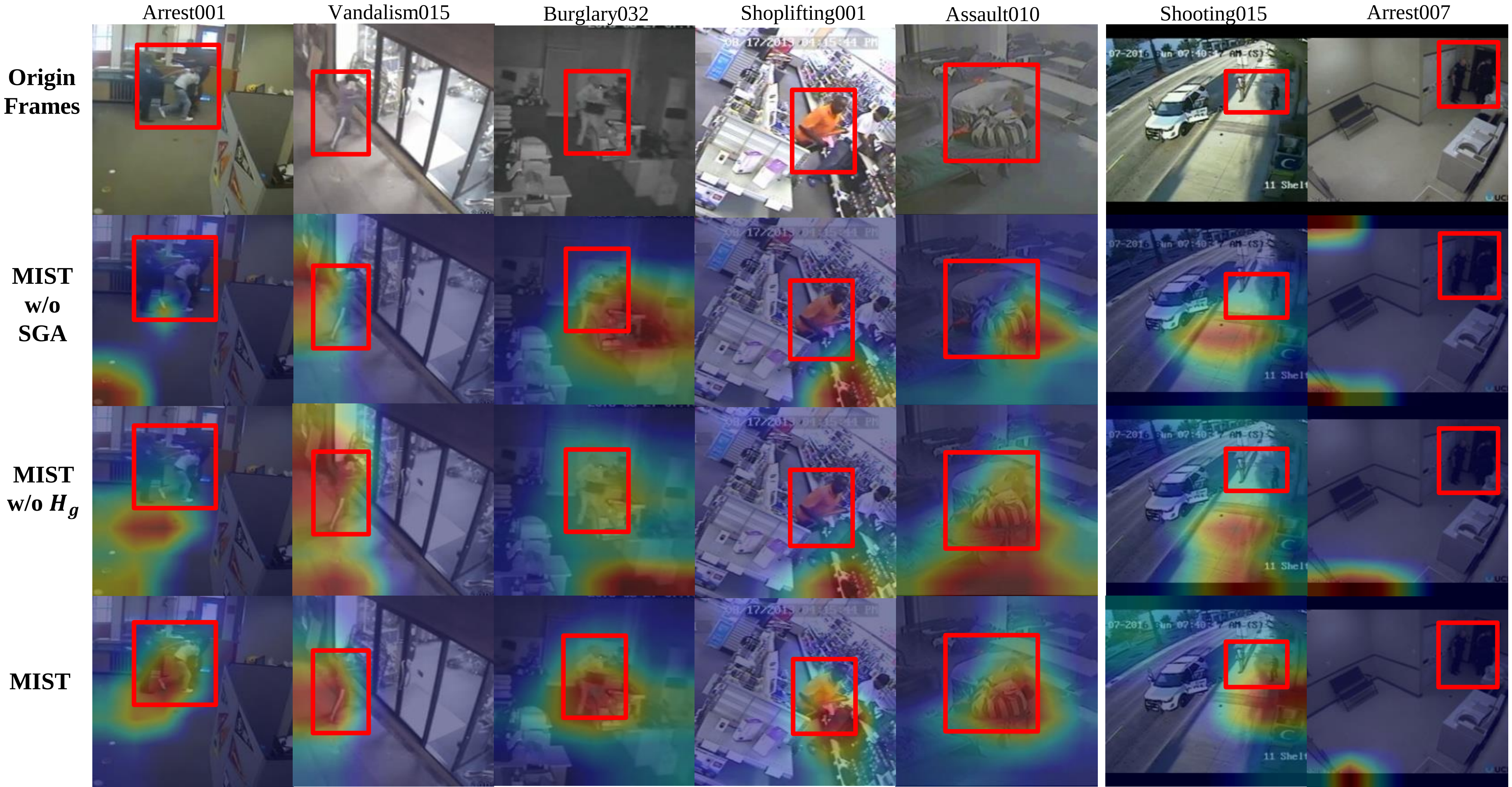
Dataset	Feature	AUC (%)		ΔAUC (%)
		Uniform	Sparse Continuous	
UCF-Crime	$C3D^{RGB}$	74.29	75.51	+1.22
	$I3D^{RGB}$	78.72	79.37	+0.65
ShanghaiTech	$C3D^{RGB}$	83.68	86.61	+2.93
	$I3D^{RGB}$	83.10	89.15	+6.05

Table 4: Performance comparisons of sparse continuous sampling and uniform sampling for MIL generator training.

Method	AUC (%)	Score Gap (%)
Baseline	74.13	0.375
$\text{MIST}^{\text{w/o PLs}}$	73.33	0.443
$\text{MIST}^{\text{w/o } H_g}$	81.97	15.37
$\text{MIST}^{\text{w/o SGA}}$	80.28	12.74
MIST	**82.30**	**17.71**

Table 5: Ablation Studies on UCF-Crime with $I3D^{RGB}$. Baseline is the original **I3D** trained with video-level labels [35]. MIST is our whole model. $\text{MIST}^{\text{w/o PLs}}$ is trained without pseudo labels but with video-level labels. $\text{MIST}^{\text{w/o } H_g}$ is MIST trained without H_g. $\text{MIST}^{\text{w/o SGA}}$ is trained without the self-guided attention module).

Spatial Explanation / Visualization on UCF-Crime



Project Page

Speed and Complexity

Project Page

Model	#Params	Speed (FPS)	FLOPs (MAC)
MIST-I3D	31M	324.46	45.68G
MIST-C3D	85M	197.10	39.26G
Zhong-C3D[35]	78M	130.04	386.2G

Table 6: Speed and computational complexity comparisons with Zhong *et al.* [35].

Demo

Project Page

Future works

- Cooperate with Label Noise Learning / Filtering
- The better alternative for two parts of MIST
- Better dataset [Clear Anomaly Definition, High Resolution…]
- …

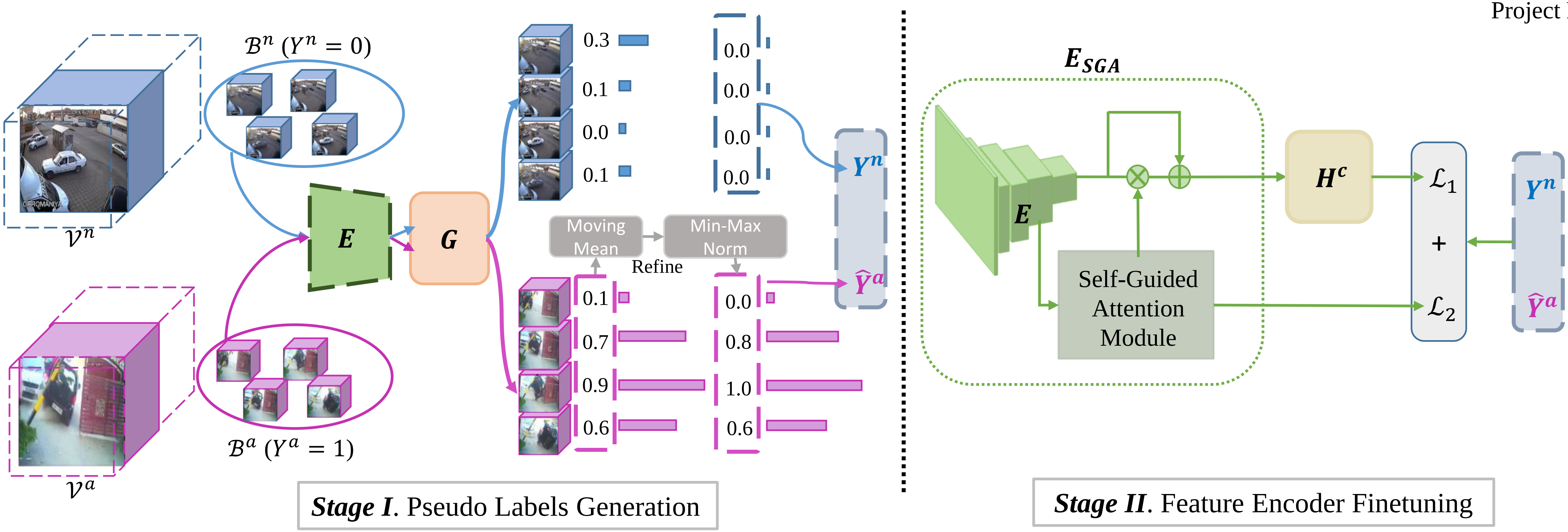
Project Page

Thanks for listening!

Project Page

https://kiwi-fung.win/2021/04/28/MIST/

Q&A



Project Page