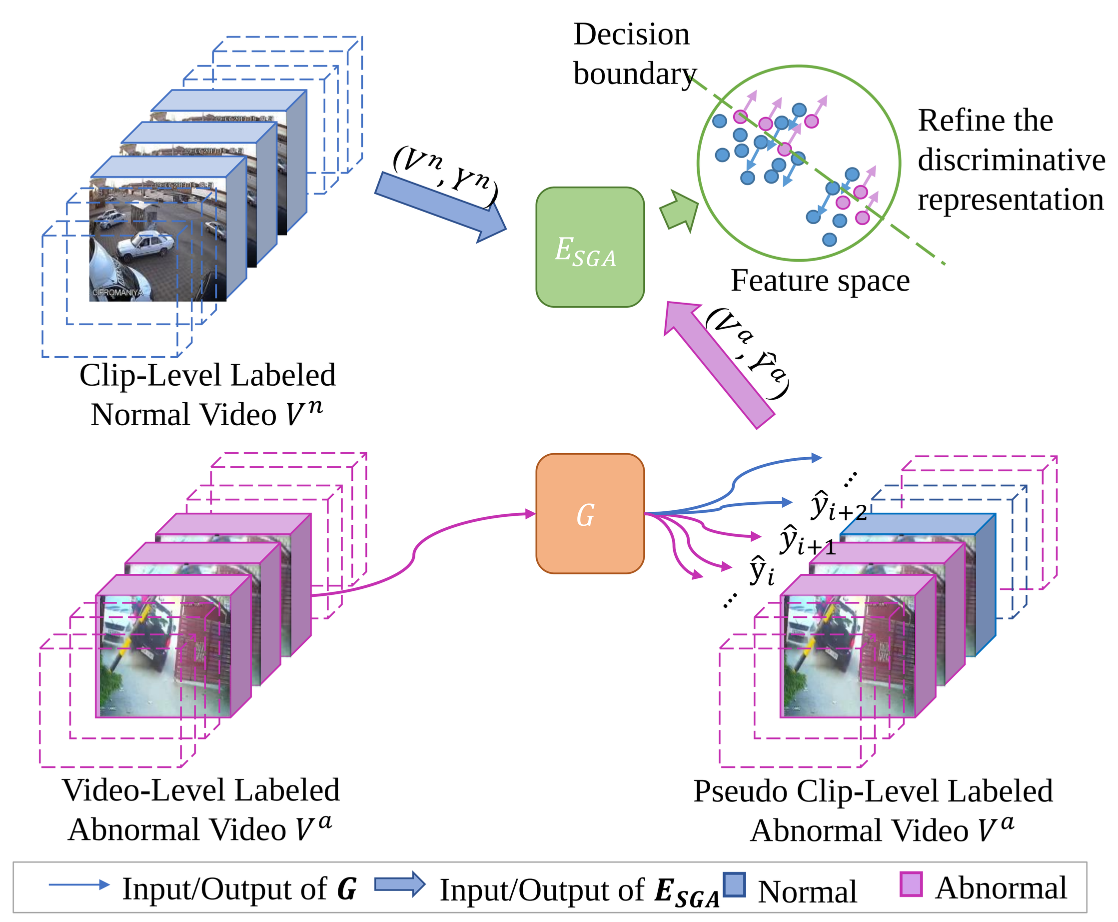
MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection

Jia-Chang Feng, Fa-Ting Hong, Wei-Shi Zheng
Sun Yat-sen University

Project page

CVPR VIRTUAL JUNE 19-25

Introduction



Introduction & Motivations

- There is a domain gap lying between common videos and surveillance videos leading to insufficient representations for video anomaly detection (VAD) that need to be minimized.
- Most previous works tackled weakly supervised VAD (WS-VAD) in coarse-grained or off-line manner that is not practical for real-time streaming videos.
- Spatial anomaly explanation/visualization is also significant for anomaly alarms understanding and solving.

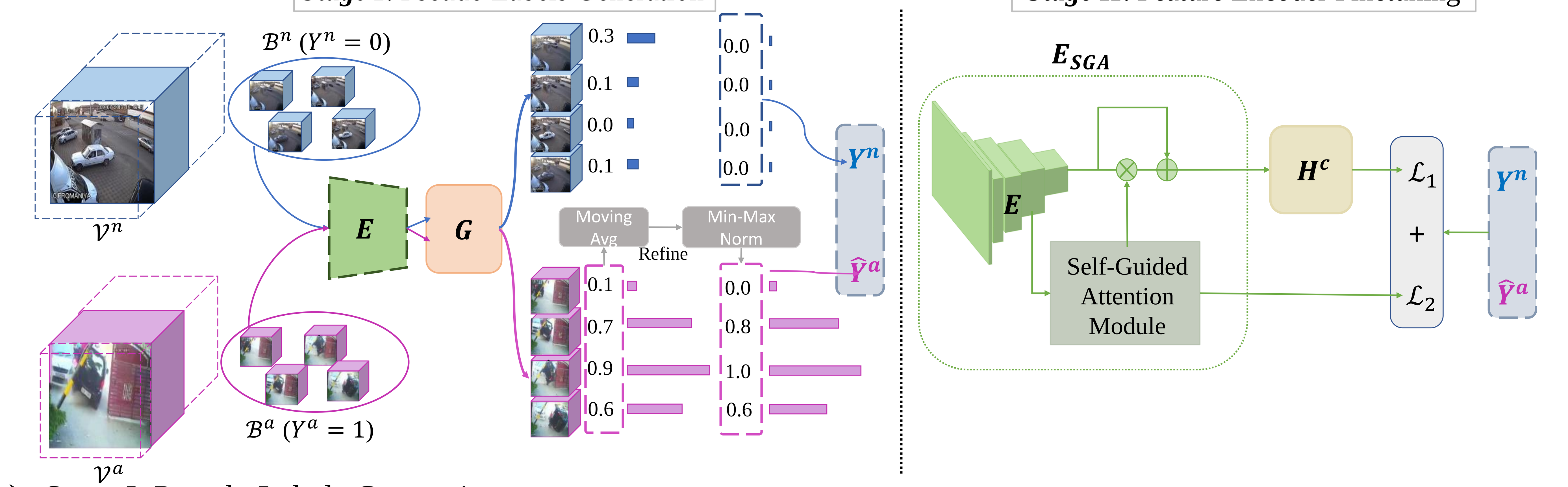
Contributions:

- The proposed two-stage framework MIST is an efficient method to finetune feature encoder for discriminative representations to tackling WS-VAD problem.
- MIST contains a multiple instance learning based pseudo label generator along with a novel sparse continuous sampling strategy, and a self-guided attention enhanced feature encoder finetuned with generated pseudo labels.
- MIST not only provide temporal anomaly detection but also provide spatial explanation/visualization.
- Extensive experiments on UCF-Crime verify the efficacy of MIST on WS-VAD.

Methodology

Overview

***Stage I*. Pseudo Labels Generation** ***Stage II*. Feature Encoder Finetuning**



Algorithm 1 Multiple instance self-training framework

Input: Clip-level labeled normal videos $V^n = \{v_i^n\}_{i=1}^N$ and corresponding clip-level labels Y^n, video-level labeled abnormal videos $V^a = \{v_i^a\}_{i=1}^N$, pretrained vanilla feature encoder E.
Output: Self-guided attention boosted feature encoder E_{SGA}, multiple instance pseudo label generator G, clip-level pseudo labels $\hat{Y}^a$ for V^a

Stage I. Pseudo Labels Generation.

1: Extract features of V^a and V^n from E as $\{\mathrm{f}_i^a\}_{i=1}^N$ and $\{\mathrm{f}_i^n\}_{i=1}^N$.
2: Training G with $\{\mathrm{f}_i^a\}_{i=1}^N$ and $\{\mathrm{f}_i^n\}_{i=1}^N$ and their corresponding video-level labels according to Eq. 7.
3: Predict clip-level pseudo labels for each clip of V^a via trained G as $\hat{Y}^a$.

Stage II. Feature Encoder Fine-tuning.

4: Combine E with self-guided attention module as E_{SGA}, then fine-tune E_{SGA} with supervision of $Y^n \cup \hat{Y}^a$.

Stage I: Pseudo Labels Generation

- Sparse continuous sampling:
  - Uniformly sample L sub-bag, where each sub-bag consists of T continuous clips.
- Generator training objective

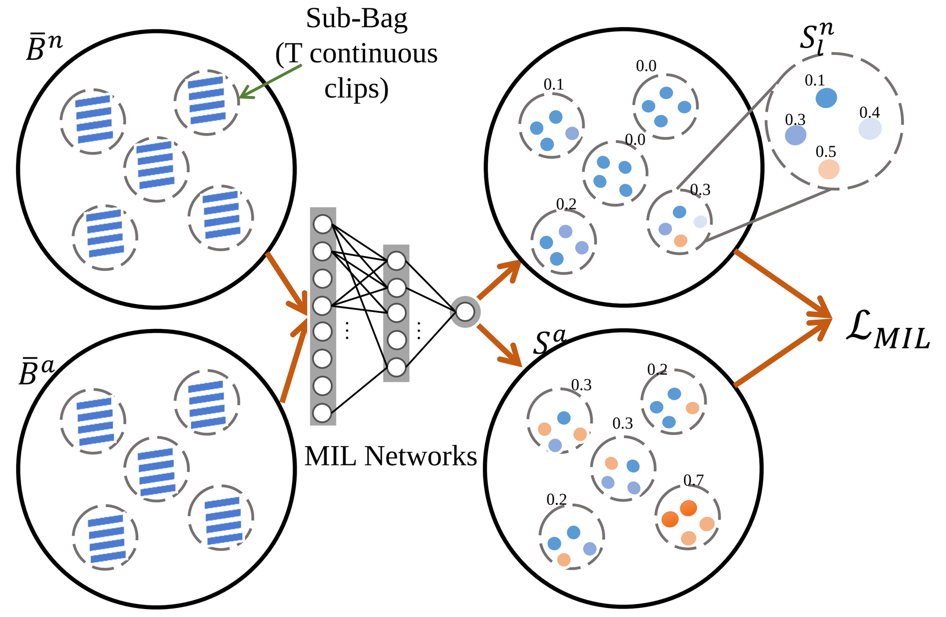
$$\mathcal{L}_{MIL} = \left(\epsilon - \max_{1 \le l \le L} \mathcal{S}_l^a + \max_{1 \le l \le L} \mathcal{S}_l^n\right)_+ + \frac{\lambda}{L}\sum_{l=1}^{L} \mathcal{S}_l^a.$$

- 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

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



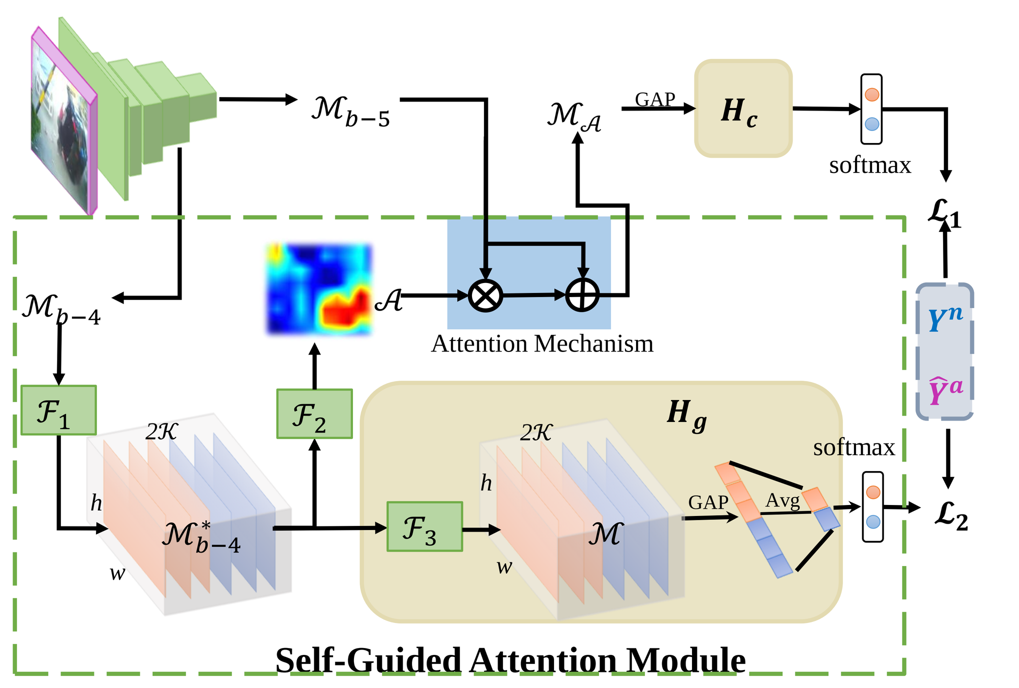
Stage II: Feature Encoder Finetuning

- Attention map generation
  - $\mathcal{A} = \mathcal{F}_2(\mathcal{F}_1(\mathcal{M}_{b-4}))$
- Attention mechanism

$\mathcal{M}_{\mathcal{A}} = \mathcal{M}_{b-5} + \mathcal{A} \circ \mathcal{M}_{b-5}$

- Attention map is indirectly enhanced by pseudo labels guidance with a guided classification head $\boldsymbol{H}_g$ to make $\mathcal{M}_{b-4}^*$ more discriminative.
- 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

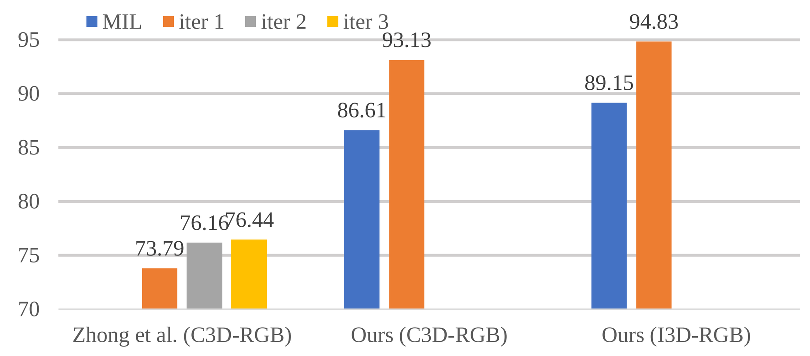
Experimental results

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.

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**

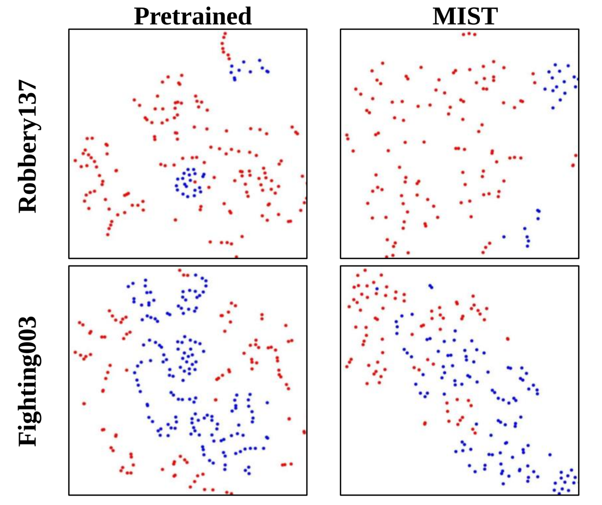
Table 2: Quantitative comparisons with existing methods on ShanghaiTech. The results with * are re-implemented.



Effect of MIST finetuning

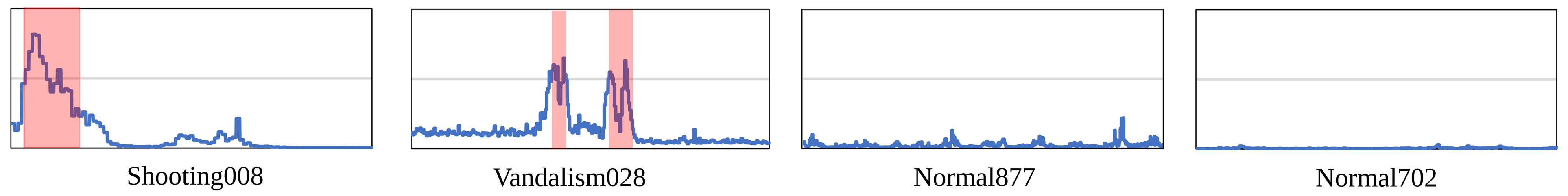
Encoder-Agnostic Methods	AUC (%) UCF-Crime pretrained	MIST	Δ	ShanghaiTech pretrained	MIST	Δ
Sultani *et al.* [20]	78.43	81.42	**+2.99**	86.92	92.63	**+5.71**
Zhang *et al.* [28]	78.11	81.58	**+3.47**	88.87	92.50	**+3.63**
AR-Net [24]	78.96	82.62	**+3.66**	85.38	92.27	**+6.89**
Our MIL generator	79.37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



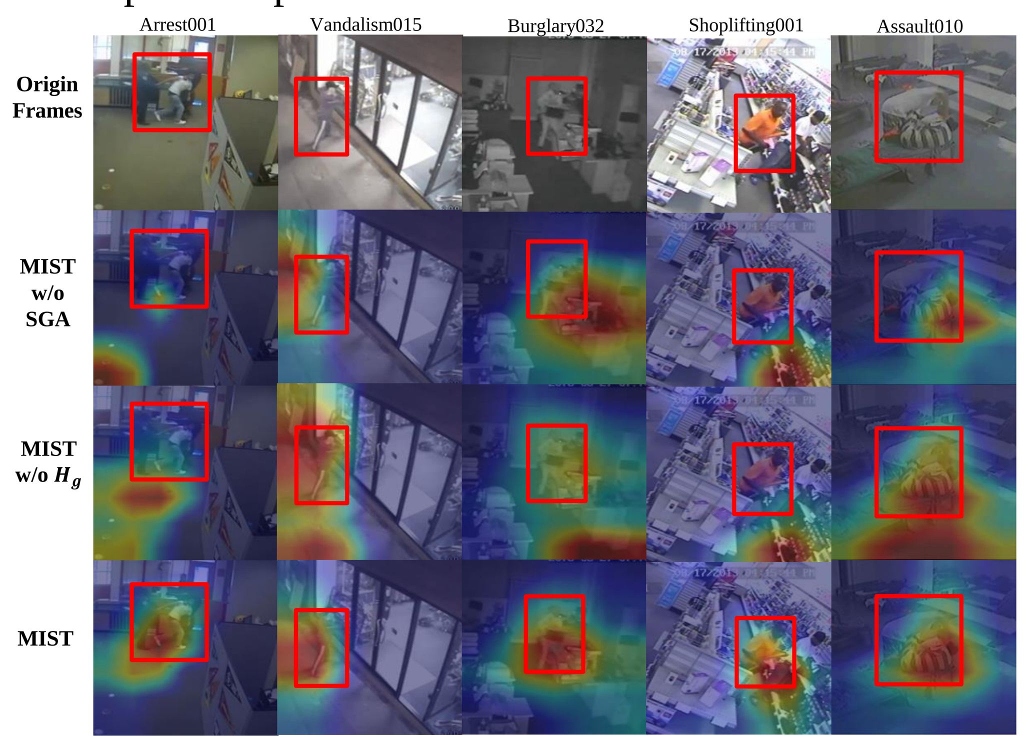
Feature space visualization via t-SNE on UCF-Crime.

Anomaly scores visualization on UCF-Crime



Ablation studies

- Spatial explanation/visualization



Dataset	Feature	AUC (%) Uniform	Sparse Continuous	ΔAUC (%)
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).