#9264

Cross-Epoch Learning for Weakly Supervised Anomaly Detection in Surveillance videos

Shenghao Yu[1], Chong Wang[1*], Qiaomei Mao[1], Yuqi Li[1] and Jiafei Wu[2]

[1]*Faculty of Electrical Engineering and Computer Science, Ningbo University, China*

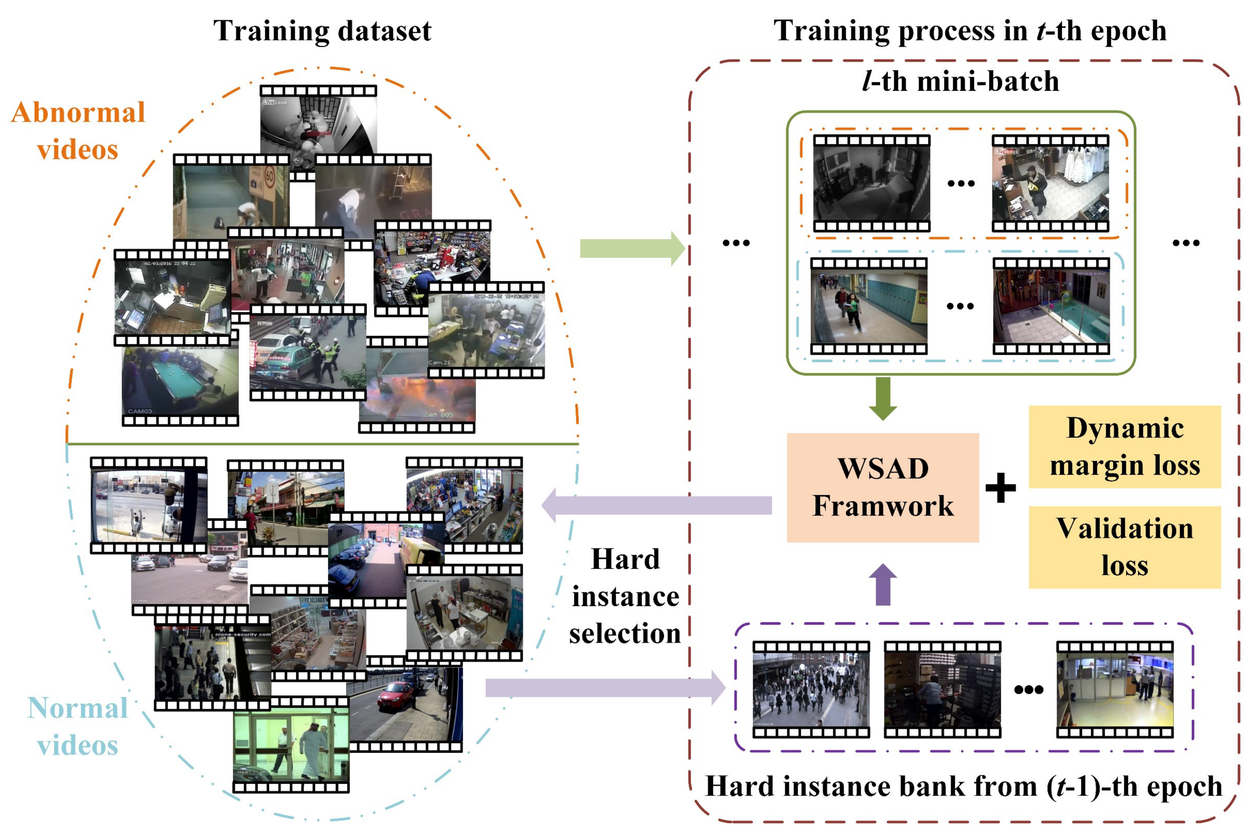
[2] *SenseTime Research, China*

icassp 2022 Singapore

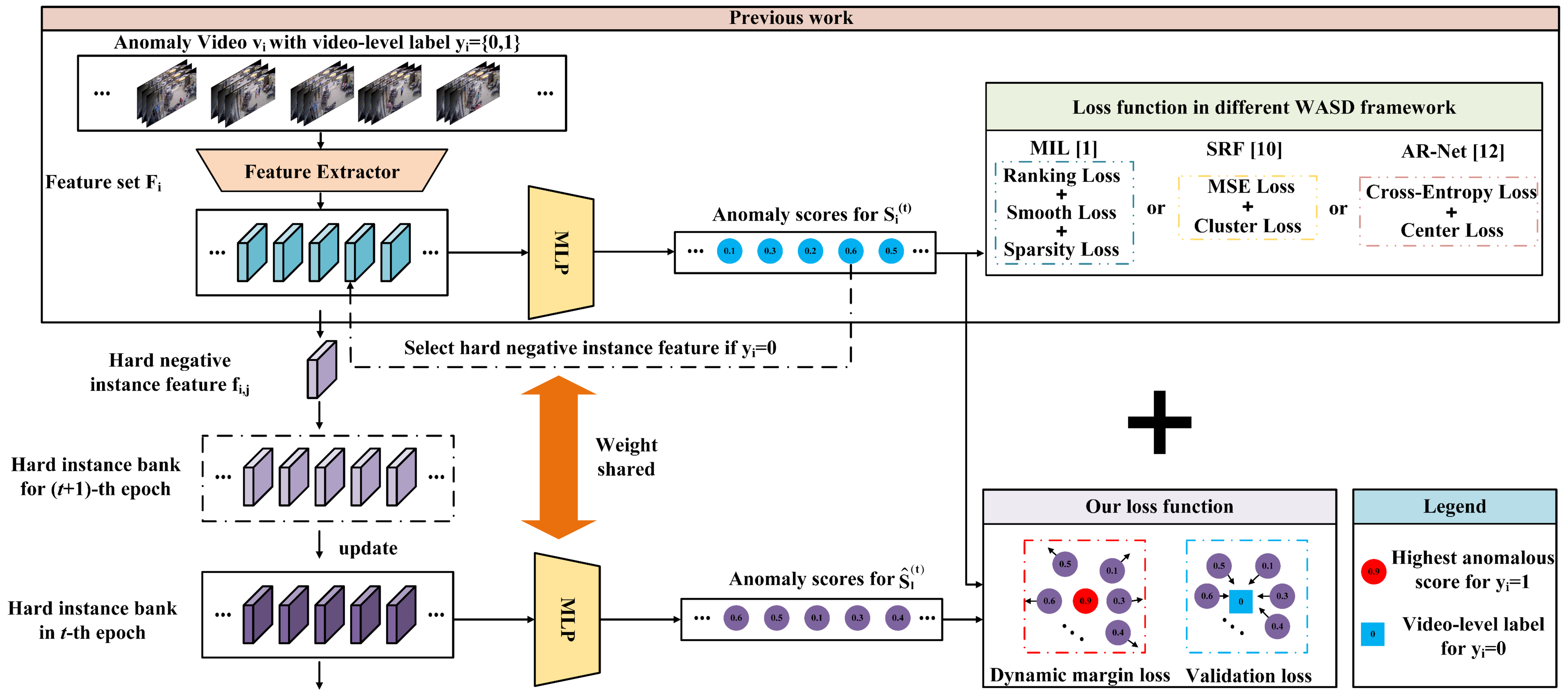
Introduction

Weakly Supervised Anomaly Detection (WSAD) in surveillance videos is a complex task since usually only video-level annotations are available. Previous work treated it as a regression problem by giving different scores on normal and anomaly events. However, the widely used mini-batch training strategy may suffer from the data imbalance between these two types of events, which limits the model's performance.

Inspired by the widely used Focal Loss in object detection, a cross-epoch learning (XEL) model is proposed to focus on the complicated cases in this paper.



Network Architecture



The overall architecture of proposed HIB embedded model. The upper part is the previous framework and the lower part is our approach. A hard instance bank (HIB) is designed to collect hard negative instances from normal events at the end of each epoch during the training stage. This HIB is utilized to a supplementary package for every mini-batches in the next epoch. Furthermore, two new losses for WSAD, namely validation loss and dynamic margin loss, are applied to not only enlarge the inter-class score distance between abnormal and normal events, but also reduce the intra-class score distance within normal events. It is worth noting that the propose XEL scheme is compatible to most previous WSAD frameworks.

Hard Instance Bank

An hard instance bank (HIB) is proposed to collect the information across multiple batches or epochs. Specifically, M hard negative instances, i.e. clip features with the highest anomaly scores in each normal video, are selected to update the HIB ($\Omega \in R^{M\times d}$) with XEL strategy.

1) Updating HIB

Considering the factor that the hardest negative instance are selected from each normal video, it is natural to update the HIB using an epoch-wise strategy. Specifically, all the clips from normal videos are re-evaluated after each training epoch. The features of those hard instances with the highest scores are picked out (e.g. t-th epoch and i-th normal video):

$$h_i^{(t)} = \underset{h_i^{(t)}\in[1,k_i]}{\operatorname{argmax}}\,(s_{i,1}^{(t)}, s_{i,2}^{(t)}, \dots, s_{i,k_i}^{(t)})$$

where $h_i^{(t)}$ is the index for the highest score in $S_i^{(t)}$. The HIB is updated at the beginning of each training epoch:

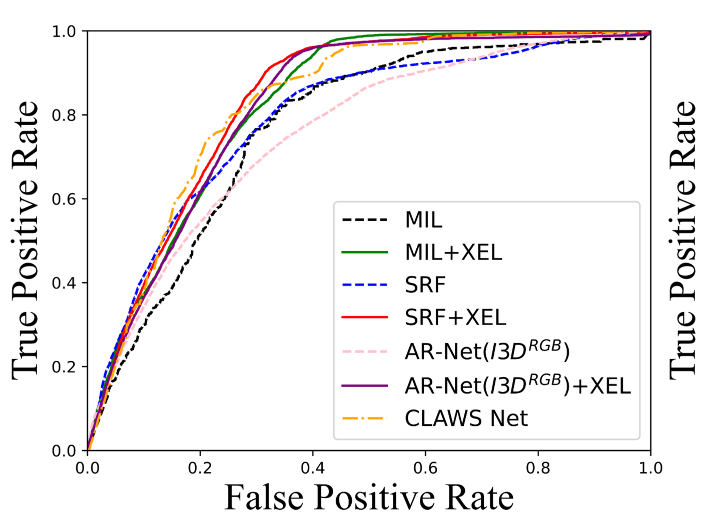
$$\Omega^{(t+1)} = \{f_{i,h_i^{(t)}}\}_{i=1}^{M}$$

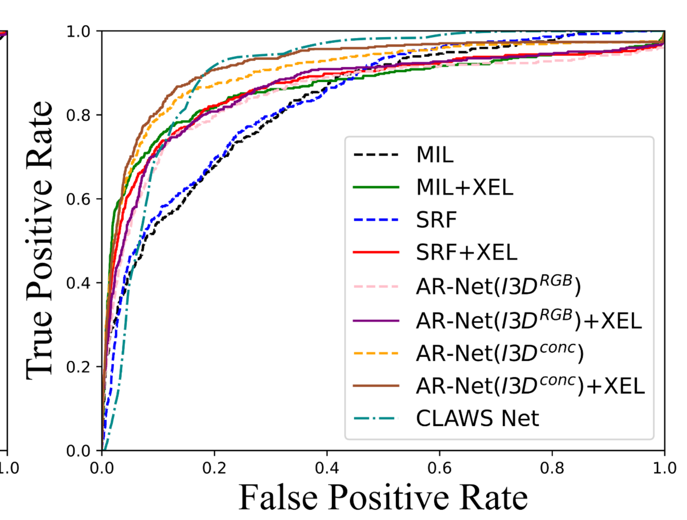
2) Learning with HIB

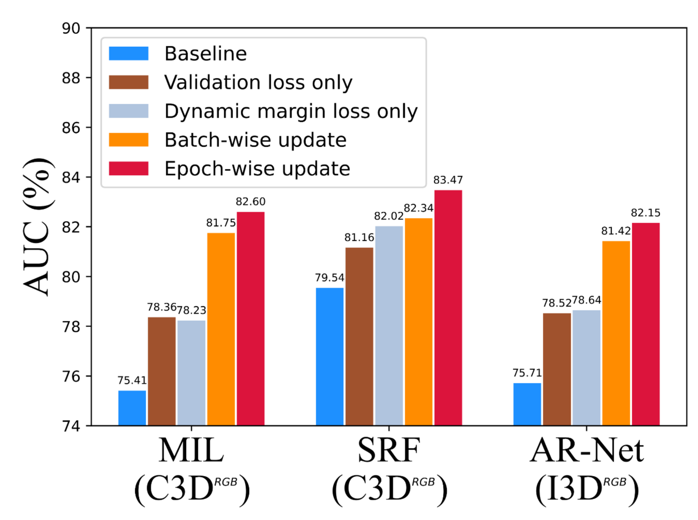
At l-th iteration in (t+1)-th epoch, the anomaly score vector $\hat{S}_l^{(t+1)}$ of the features in HIB are calculated in every iteration:

$$\hat{S}_l^{(t+1)} = \{\hat{s}_{i,h_i^{(t)},l}^{(t+1)}\}_{i=1}^{M} = \{\mathrm{MLP}_l^{(t+1)}(f_{i,h_i^{(t)}})\}_{i=1}^{M}$$

Experimental Results







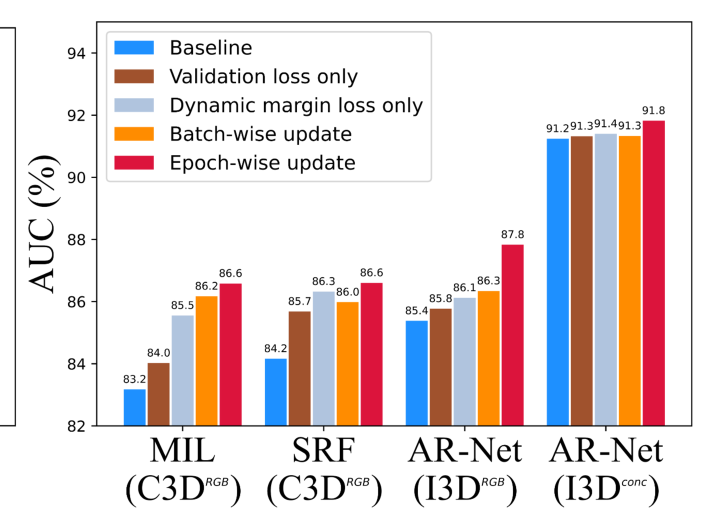


TABLE II
FALSE ALARM RATE (%) AND TRUE POSITIVE RATE COMPARISON ON NORMAL TEST VIDEOS ON UCF-CRIME DATASET

Method	Feature type	False Alarm Rate (%)	True Positive Rate
SVM Baseline	C3D	-	-
Hasan et al. [22]	C3D	27.2	-
Lu et al. [23]	C3D	3.1	-
MIL [1]	C3D	1.9	0.21
Zhong et al. [9]	C3D	2.8	-
Zhong et al. [9]	TSN^{RGB}	1.1	-
SRF [10]	C3D	0.13	0.25
ClAWS Net [11]	C3D	0.12	-
AR-Net [12]	$I3D^{RGB}$	0.40	0.13
MIL+XEL	C3D	**0.0** (↓ 1.9)	0.44
SRF+XEL	C3D	**0.0** (↓ 0.13)	**0.45**
AR-Net+XEL	$I3D^{RGB}$	0.03 (↓ 0.37)	0.40

TABLE I
FRAME-LEVEL AUC (%) PERFORMANCE COMPARISON

Method	Feature type	UCF-Crime	ShanghaiTech
SVM Baseline	C3D	50.00	-
Hasan et al. [22]	C3D	50.60	-
Lu et al. [23]	C3D	65.51	-
MIL [1]	C3D	75.41	83.17*
Zhong et al. [9]	C3D	81.08	76.44
Zhong et al. [9]	TSN^{RGB}	82.12	84.13
SRF [10]	C3D	79.54	84.16
ClAWS Net [11]	C3D	83.08	89.67
AR-Net [12]	$I3D^{RGB}$	75.71*	85.38
AR-Net [12]	$I3D^{conc}$	-	91.24
MIL+XEL	C3D	82.60	86.58
SRF+XEL	C3D	**83.47**	86.60
AR-Net+XEL	$I3D^{RGB}$	82.15	87.83
AR-Net+XEL	$I3D^{conc}$	-	**91.82**

* indicate we re-implement the framework in our experiments.

The experiments were conducted on two datasets, including ShangahiTech and UCF-Crime. The effectiveness of XEL is shown by the Receiver Operating Characteristic (ROC) curves, corresponding area under the curve (AUC) and false alarm rate (FAR). The re-implemented frameworks generally have better performance at various thresholds of ROC. All three XEL embedded frameworks (MIL, SRF and AR-Net) achieve better AUC than their vanilla forms with noticeable improvement (7.19%, 3.93%, 6.44 % on UCF-Crime, and 3.41%, 2.44%, 2.45% on ShanghaiTech dataset). In the case of UCF-Crime, the performance of all three frameworks are boosted about 2% by each loss function in the proposed XEL. Similar trends also shown in experiments on ShanghaiTech datasets. Meanwhile, the AUCs of batch-wise updating strategy are constantly lower than the epoch-wise updating strategy.

Loss Function

A validation loss L_v is defined to penalize the hard instance:

$$L_v = \frac{1}{M}\sum_{i=1}^{M} |\hat{s}_{i,h_i^{(t)},l}^{(t+1)} - y_{i,h_i^{(t)}}|$$

A dynamic margin loss L_m is proposed with a maximum margin ε between the hard negative instances in HIB and the most abnormal instances in abnormal videos:

$$L_m = \frac{1}{M}\sum_{i=1}^{M} \max(0, \varepsilon - \max\left(S_a^{(t+1)}\right) + \hat{s}_{i,h_i^t,l}^{(t+1)})$$

The final loss is defined as:

$$L = L_o + \lambda_1 L_v + \lambda_2 L_m$$

L_o is the loss function of any given WSAD framework.