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

Shenghao Yu, Chong Wang, *Member, IEEE*, Qiaomei Mao, Yuqi Li, and Jiafei Wu

***Abstract*—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. In this work, a cross-epoch learning (XEL) strategy associated with a hard instance bank (HIB) is proposed to introduce additional information from previous training epochs. Two new losses are proposed for XEL to achieve a higher detection rate as well as a lower false alarm rate of anomaly events. Moreover, the proposed XEL can be directly integrated into any existing WSAD framework. Experimental results of three XEL embedded models have shown promising AUC improvement (3%∼7%) on two public datasets, surpassing the state-of-the-art methods. Our code is available at: https://github.com/sdjsngs/XEL-WSAD.**

***Index Terms*—Anomaly detection, weakly supervised learning, cross-epoch learning, video understanding.**

I. Introduction

In recent years, surveillance cameras have been widely used in public places. An important task of surveillance system is to detect anomalous events [1]–[3], e.g., deliberate arson, fights, explosions, etc. Attributed to the low incidence of abnormal events in real life, anomalies are usually regarded as the appearance or behavior that differs from the previous normal events. Therefore, unsupervised models are utilize by some previous works [4]–[8] to train a one-class classifier for learning the general patterns of normal events. Then, those events with unseen patterns will be recognized as abnormal ones. However, it is impractical to collect all normal behaviors. Thus, any normal events deviated from the prior encoded modes may be false alarmed.

To address this issue, anomaly detection is considered as a binary classification problem [9]–[11] in which the training data contains both anomalous videos and normal videos. Training a binary classification model for anomaly detection is a weakly supervised learning problem, since only the untrimmed videos and video-level labels are provided. There are three steps in the general framework of Weakly Supervised Anomaly Detection (WSAD), 1) the videos are divided into clips and converted to spatio-temporal features by a pre-trained feature extractor, 2) the anomaly scores are generated by a Multilayer Perceptron (MLP) and 3) the model is optimized by the loss function. The WSAD is formulated as a multi-instance learning (MIL) task in Sultain *et al.* [1]. Each video is treated as a bag and divided into multiple non-overlapping clips as instances. A ranking loss is used to widen the gap between the two highest-scoring instances in the positive and negative bags, respectively. By replacing the max selection with a k-max selection method in the MIL framework, a better performance is achieves by AR-Net [12]. Meanwhile, WSAD is treated as an action recognition task with noisy labels in Zhang *et al.* [9]. An action classifier is trained with noise labels, while another GCN is trained as the noise cleaner. Instead of cleaning the noisy label, a binary clustering based self-reasoning framework [10] (SRF) is proposed to generate pseudo label for WSAD.

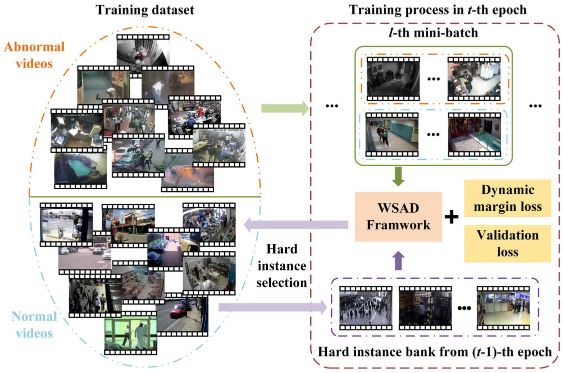

Fig. 1. The overview of the proposed cross-epoch learning.

Since the data is highly imbalanced in WSAD, introducing extra information from previous min-batches will help the model training. However, the patterns of normal events are extremely complex to model. Simply increasing the data amount of each mini-batch, such as cross-batch memory (XBM) [13] or long-term feature banks [14], is not as effective as in deep metric learning or video understanding.

Inspired by the widely used Focal Loss in object detection [15], [16], a cross-epoch learning (XEL) model is proposed to focus on the complicated cases in this paper as shown in Fig. 1. To be specific, a hard instance bank (HIB) is designed to collect hard negative instances from normal events at the end of

Manuscript received June 29, 2021; revised September 20, 2021; accepted September 28, 2021. Date of publication October 6, 2021; date of current version November 4, 2021. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY20F030005 and in part by the National Natural Science Foundation of China under Grant 61603202. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Parvaneh Saeedi. *(Corresponding author: Chong Wang.)*

Shenghao Yu, Chong Wang, Qiaomei Mao, and Yuqi Li are with the Faculty of EECS, Ningbo University, Ningbo 315210, China (e-mail: ysh_nbu@163.com; wangchong@nbu.edu.cn; qiao_ali@foxmail.com; liyuqi1@nbu.edu.cn).

Jiafei Wu is with SenseTime Research, Shanghai 200233, China (e-mail: wujiafei@sensetime.com).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LSP.2021.3117737, provided by the authors.

Digital Object Identifier 10.1109/LSP.2021.3117737

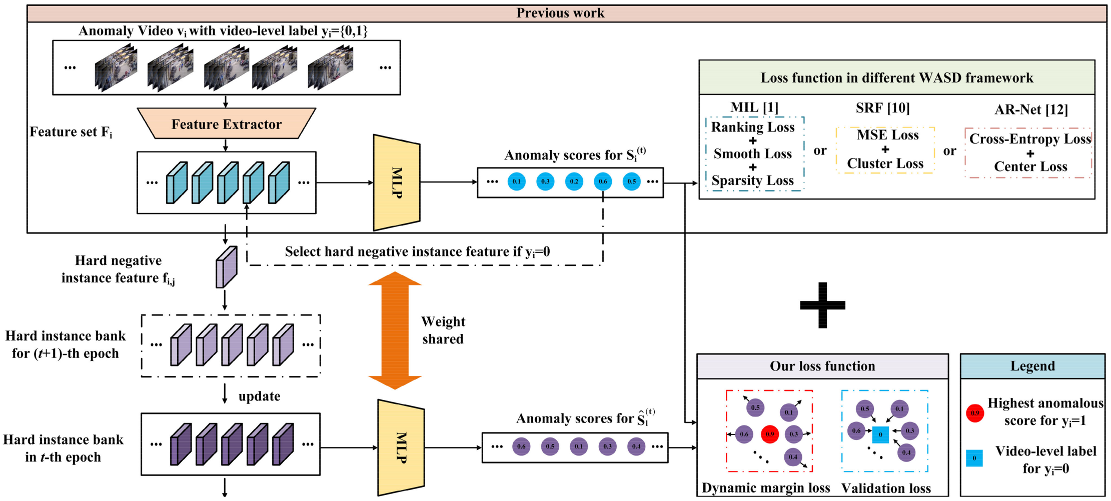
Previous work
Anomaly Video v_i with video-level label y_i={0,1}
Feature set F_i
Feature Extractor
MLP
Anomaly scores for $S_i^{(t)}$
0.1 0.3 0.2 0.6 0.5
Loss function in different WASD framework
MIL [1]
Ranking Loss + Smooth Loss + Sparsity Loss
or
SRF [10]
MSE Loss + Cluster Loss
or
AR-Net [12]
Cross-Entropy Loss + Center Loss
Hard negative instance feature $f_{i,j}$
Select hard negative instance feature if y_i=0
Weight shared
Hard instance bank for (t+1)-th epoch
update
Hard instance bank in t-th epoch
MLP
Anomaly scores for $\hat{S}_i^{(t)}$
0.6 0.5 0.1 0.3 0.4
+
Our loss function
Dynamic margin loss
Validation loss
Legend
Highest anomalous score for y_i=1
Video-level label for y_i=0

Fig. 2. The overall architecture of proposed HIB embedded model. The upper part is the previous framework, the video is divided into clips and converted to spatio-temporal features. The MLP is used to calculate the anomaly scores. The lower part is our approach. The hard instances in each epoch are selected and collected to make up our HIB and used to optimize the MLP with our extra two loss functions.

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.

The rest of the paper is organized as follows. First, the framework of WSAD is briefly reviewed in Section II-A. The proposed XEL and HIB is then described in Section II-B. The performance of three XEL embedded WSAD frameworks is evaluated with extensive experiments in Section III. Finally, this paper is concluded in Section IV.

II. Methodology

An anomaly detection dataset consists of N untrimmed videos $V = \{v_i\}_{i=1}^{N}$ associated with video-level labels $Y = \{y_i\}_{i=1}^{N}$, where $y_i = \{0, 1\}$ (0: normal; 1: abnormal). The number of normal videos is M. The goal of WSAD is to detect the abnormal events once they occur in the test videos. Specifically, the i-th anomalous video in V would be divided into k_i clips and the corresponding anomaly score vector in t-th epoch can be viewed as:

$$S_{a,i}^{(t)} = \left\{ s_{a,i,j}^{(t)} \right\}_{j=1}^{k_i}, \tag{1}$$

where $s_{a,i,j}^{(t)} \in [0, 1]$ is the j-th anomaly score in $S_{a,i}^{(t)}$ and a higher score value indicates a higher probability of abnormality.

The overall framework of this paper is shown in Fig. 2. Firstly, the input video is divided into multiple clips. Then a feature extractor is applied to obtain the spatio-temporal features, and the anomaly scores for all clips are generated by a Multi-Layer Perceptron (MLP). Those ones higher than a given threshold will be considered as anomaly events. Among these processes, a hard instance bank (HIB) is plugged in to utilize those hard negative instances, i.e., ones with high scores but from normal videos, to perform a cross-epoch learning (XEL). It will prevent normal frames from being incorrectly detected as anomaly instances

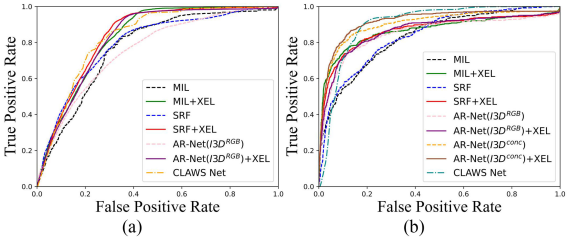

Fig. 3. ROC curve of different frameworks on (a) UCF-Crime and (b) ShanghaiTech. Best viewed in color.

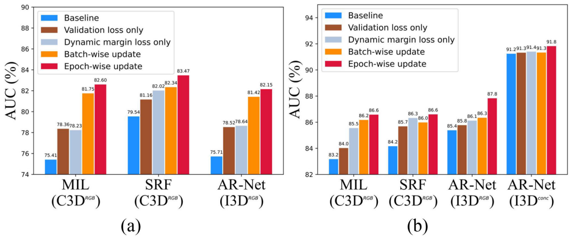

Fig. 4. Ablation studies on (a) UCF-Crime and (b) ShanghaiTech.

A. General Framework of WSAD

Video Split: Following the above definition, input video v_i is divided into k_i non-overlapping temporal clips in the first step of WSAD,

$$v_i = \{c_{i,j}\}_{j=1}^{k_i}, \tag{2}$$

where $c_{i,j}$ is the j-th clip in v_i.

Feature Extractor: A pre-trained neural network is then applied to extract the spatio-temporal features from the video clips. Two widely used networks in WSAD, namely Convolution 3D (C3D) [17] and Inflated 3D (I3D) [18], are tested in our experiments. The video v_i is thus transferred into a feature set F_i,

$$F_i = \{f_{i,j}\}_{j=1}^{k_i} = \{g(c_{i,j})\}_{j=1}^{k_i}, \tag{3}$$

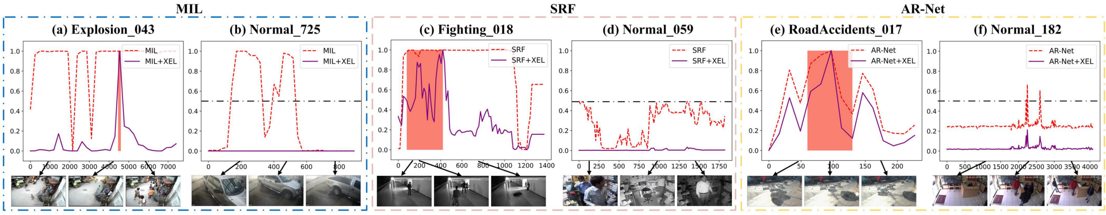

Fig. 5. Examples of the proposed HIB models on UCF-Crime test videos. Red regions are the anomalous ground truth. (a), (c) and (e) show the predicted scores on anomalous videos, while (b), (d) and (f) show the scores on normal videos. The black dash-dot lines are the false alarm rate threshold (0.5) for normal videos.

where $f_{i,j}$ is the *j*-th feature in F_i, d is the feature dimension and *g(*)* is the feature extractor.

Anomaly Score Generator: A Multi-Layer Perceptron (MLP) is utilized to generate anomaly scores. Generally, 2-3 layers are enough. With the input $f_{i,j}$, the output of MLP in *t*-th epoch is the desired anomaly score $s_{i,j}^{(t)}$:

$$s_{i,j}^{(t)} = \mathrm{MLP}^{(t)}\ (f_{i,j})\,. \tag{4}$$

Loss Function: The goal of the loss function in WSAD is to guide the model to learn the discriminative ability for normal and abnormal events. Besides the conventional loss functions used in previous works [1], [10], [12], two novel loss functions, namely validation loss and dynamic margin loss, is proposed to exploit the diversity of dataset by using the HIB.

B. Hard Instance Bank

In this paper, an 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. Served as the boundary between the normal and abnormal events, such HIB associated with two new losses have two merits, 1) the scores of these hard negative instances are suppressed using a validation loss to achieve a lower false alarm rate; 2) the lower bound of the highest scores in anomalous videos is lifted higher with a dynamic margin loss to boost the detection rate. During the training process, the HIB is utilized to validate the MLP every iteration at slight cost on RAM, GPU memory and running time.

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):

$$h_i^{(t)} = \underset{h_i^{(t)}\in[1,k_i]}{\mathrm{argmax}}\left(s_{i,1}^{(t)}, s_{i,2}^{(t)}, \ldots, s_{i,k_i}^{(t)}\right), \tag{5}$$

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:

$$\Omega^{(t+1)} = \left\{f_{i,h_i^{(t)}}\right\}_{i=1}^{M}. \tag{6}$$

It is worth noting that an alternate updating strategy of HIB is to maintain it as a dynamic queue used in [16], i.e., the newest features of current mini-batch are used to replace the oldest features in the memory bank. One drawback of such strategy is the large oscillation cost for the memory bank [19]–[21]. Moreover, its performance is also inferior to the proposed epoch-wise updating strategy as shown in Section III.

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:

$$\hat{S}_l^{(t+1)} = \left\{\hat{s}^{(t+1)}_{i,h_i^{(t)},l}\right\}_{i=1}^{M} = \left\{\mathrm{MLP}_l^{(t+1)}\left(f_{i,h_i^{(t)}}\right)\right\}_{i=1}^{M}, \tag{7}$$

This can be viewed as a validation process of currently learned $\mathrm{MLP}_l^{(t+1)}$ to evaluate its performance on normal videos.

Obviously, the anomaly scores of normal videos should be as close to zero as possible. Thus, a validation loss is defined as to penalize the hard negative instances,

$$L_v = \frac{1}{M}\sum_{i\,=\,1}^{M}\left|\hat{s}^{(t+1)}_{i,h_i^{(t)},l} - y_{i,h_i^{(t)}}\right|, \tag{8}$$

where $y_{i,h_i^{(t)}}$ is the clip-level label for the instance feature $f_{i,h_i^{(t)}}$. By introducing the extra hard instances from the previous epoch, the performance for normal events is less likely to be degraded by the other losses or data in the certain mini-batch.

Meanwhile, a dynamic margin loss function is proposed with a maximum margin ε between the hard negative instances in HIB and the most abnormal instances in abnormal videos,

$$L_m = \frac{1}{M}\sum_{i\,=\,1}^{M}\max\left(0, \varepsilon - \max\left(S_a^{(t+1)}\right) + \hat{s}^{(t+1)}_{i,h_i^t,l}\right), \tag{9}$$

where $S_a^{(t+1)}$ is the anomaly score vector for the anomalous video in (*t*+1)-th epoch. It is noting that the final goal of WSAD is to set large score to all anomalous clips. The MLP may be guided by a too large margin (such as 1) in the loss to focus only on the most distinguishable one. However, a suitable margin value is hard to determine. The margin ε is thus progressively increased during the training process in our experiments.

3) Final Loss Function: In the end, our final loss function is defined as:

$$L\ = L_o\ + \lambda_1 L_v + \lambda_2 L_m, \tag{10}$$

where $\lambda_1$and λ_2 are the weights for validation loss and dynamic margin loss, L_o is the loss function of any given WSAD framework. In our experiments, three state-of-the-art frameworks are implemented, namely MIL [1], SRF [10] and AR-Net [12]. The detailed formulations of their loss functions are summarized in the supplemental materials.

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.

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

III. EXPERIMENTS

A. Dataset

UCF-Crime [1] is a large-scale complex dataset for anomaly detection. It contains 13 real-world anomalous behaviors, distributed in 1,900 untrimmed videos with a total duration of 128 hours. Following the protocol in [1], the training split has 800 normal videos and 810 anomalous videos with video-level labels, while the test split has 150 normal videos and 140 anomalous videos with frame-level annotations.

ShanghaiTech [3] is another medium-scale dataset, including a total of 437 videos (330 normal videos and 107 anomalous videos) from 13 different cameras in the university campus. As the same as in [9], the training and test splits have 175 and 155 normal videos, 63 and 44 anomalous videos, respectively.

B. Implement Details

Three state-of-the-art anomaly detection networks, namely MIL [1], SRF [10] and AR-Net [12], have been re-implemented by embedding the proposed XEL. For fair comparison, the settings have followed their original papers and most of the hyperparameters remain the same. The batch-size for MIL, SRF and AR-Net are 1920, 1000 and 1920 (clips), respectively. Roughly the half of clips are from normal videos, while the other half are from anomaly videos. C3D is chosen as the feature extractors for MIL and SRF, while I3D is for AR-Net. Meanwhile, the size of HIB is 800 and 175 for the UCF-Crime and ShanghaiTech dataset, respectively. Adam algorithm is used as the optimizer and the initial learning rate is set as 10^{-3}, 5×10^{-5} and 10^{-4} for MIL, SRF and AR-Net, respectively. Both λ_1 and λ_2 in (10) are set to 1 to provide a same scale range. Meanwhile, the value of dynamic margin ε is gradually increased from 0.5 to 1 during the training process.

C. Experimental Results

The effectiveness of XEL is shown by the Receiver Operating Characteristic (ROC) curves, corresponding area under the curve (AUC) and false alarm rate in Fig. 3, Table I and Table II, respectively. 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).

Ablation studies on the validation loss, dynamic margin loss and the updating strategy are demonstrated in Fig. 4. 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, which indicates the epoch-wise strategy can better exploit the global distribution among the dataset.

D. False Alarm Rate and True Positive Rate

In real world applications, the false alarm rate should be as low as possible for a reliable anomaly detection model. Like [1], the false alarm rate (FAR) and true positive rate (TPR) is evaluated on UCF-Crime dataset with a threshold of 0.5. As shown in Table II, the lowest FAR and highest TPR are achieved by XEL embedded models. For those models without HIB, the normal frames are more likely to be marked as anomaly ones, which raises false alarm rate.

E. Qualitative Results and Analysis

The visual examples of anomaly score plots generated by the proposed XEL are presented in Fig. 5. One anomalous video and one normal video for each XEL embedded models (MIL, SRF and AR-Net) are provided. In the case of explosion event (Fig. 5(a)), it happened very briefly. Thus the vanilla MIL cannot accurately discriminate the normal and anomalous regions. In the case of fighting event (Fig. 5(c)), due to the dim light, SRF still gives a high anomaly score even when the fight is over. In the normal videos (Fig. 5(b, d, f)), the proposed HIB is capable to suppress the scores lower (close to 0), making it easier to discriminate normal and abnormal events.

IV. CONCLUSION

In this work, a simple yet effective cross-epoch learning (XEL) strategy is presented for WSAD in surveillance videos. To be specific, a hard instance bank (HIB) is used to collect the hard instances from normal videos and it is updated in an epoch-wise manner, which allow the model to make full use of the whole dataset. The proposed XEL can be seamlessly integrated into any conventional WSAD frameworks and improve the detection ability. In particular, three XEL embedded WSAD frameworks achieved a higher AUC on UCF-Crime and ShanghaiTech dataset.

REFERENCES

[1] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6479–6488.

[2] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Process. Lett.*, vol. 22, no. 9, pp. 1477–1481, Sep. 2015.

[3] K. Doshi and Y. Yilmaz, "Fast unsupervised anomaly detection in traffic videos," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2020, pp. 624–625.

[4] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 341–349.

[5] T.-N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2019, pp. 1273–1283.

[6] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3379–3388.

[7] D. Gong *et al.*, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE Int. Conf. Comput. Vision*, Oct. 2019, pp. 1705–1714.

[8] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 733–742.

[9] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1237–1246.

[10] M. Z. Zaheer, A. Mahmood, H. Shin, and S. I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Process. Lett.*, vol. 27, pp. 1705–1709, 2020.

[11] M. Z. Zaheer *et al.*, "CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *Proc. ECCV*, 2020, pp. 358–376.

[12] B. Wan, Y. Fang, X. Xia, and J. Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2020, pp. 1–6.

[13] X. Wang, H. Zhang, W. Huang, W. Huang, and M. R. Scott, "Cross-batch memory for embedding learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 6388–6397.

[14] C. Y. Wu *et al.*, "Long-term feature banks for detailed video understanding," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 284–293.

[15] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2980–2988.

[16] Q. Mao, C. Wang, S. Yu, Y. Zheng, and Y. Li, "Zero-shot object detection with attributes-based category similarity," *IEEE Trans. Circuits Syst. II: Exp. Briefs*, vol. 67, no. 5, pp. 921–925, May 2020.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 4489–4497.

[18] C. Joao and Z. Andrew, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6299–6308.

[19] Z. Wu, Y. Xiong, and S.X. Yu, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3733–3742.

[20] L. Suichan, C. Dapeng, L. Bin, Y. Nenghai, and Z. Rui, "Memory-based neighbourhood embedding for visual recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 6102–6111.

[21] Z. Zhong, L. Zheng, and Z. Luo, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 598–607.

[22] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 733–742.

[23] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vision*, 2013, pp. 2720–2727.