MULTI-SCALE CONTINUITY-AWARE REFINEMENT NETWORK FOR WEAKLY SUPERVISED VIDEO ANOMALY DETECTION

Yiling Gong[1], Chong Wang[1,2,], Xinmiao Dai[1], Shenghao Yu[1], Lehong Xiang[2], Jiafei Wu[3]*

[1]Faculty of Electrical Engineering and Computer Science, Ningbo University, China
[2]Loctek Ergonomic Technology Corporation, China
[3]SenseTime Research, China

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9860012

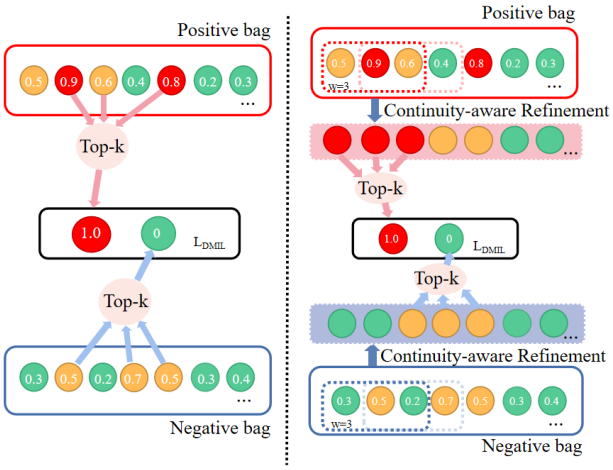
ABSTRACT

In many previous work, weakly supervised video anomaly detection is formulated as a multiple instance learning (MIL) problem, which represents the video as a bag of multiple instances. However, most MIL-based frameworks only focused on identifying anomalous events from the given instances, without considering the event continuity. Motivated by the fact that abnormal events tend to be more continuous in real-world videos, a Multi-scale Continuity-aware Refinement Network (MCR) is proposed in this paper. It utilizes the property of multi-scale continuity to refine anomaly scores by introducing differential contextual information of instances. At the same time, multi-scale attention is designed to produce a video-level weights in order to select the proper scale and fuse all scores at different scales. Experimental results of MCR show noticeable improvement on two public datasets, specifically obtaining a frame-level AUC 94.92% on ShanghaiTech dataset.

Index Terms—Anomaly detection, multiple instance learning, continuity, multi-scale attention

1. INTRODUCTION

Due to the wide range of real-world applications in surveillance, video anomaly detection has always been a potential research direction in computer vision. The goal of video anomaly detection is to identify the time window when an anomalous event happens [1]. However, different from image classification tasks, it would be more difficult and expensive to acquire large-scale frame-level annotation in videos. Since most surveillance videos are normal, some previous works [2][3][4][5] treat anomaly detection as an unsupervised learning task. They solve the problem by training a one-class classifier exclusively with normal videos. Unsupervised methods intend to learn the general patterns of normal events in the training stage and recognize those unseen patterns as anomalies in the testing stage. However, it would be unrealistic to cover all normal events in training set and those normal events that never appeared in the training set may tend to be false alarmed under this paradigm.



(a) AR-Net [8] (b) Ours

Fig. 1. Different anomaly score processing procedures compared with conventional methods. Numbers in colored circles stand for anomaly scores of clips.

Therefore, some researchers followed the binary classification paradigm, considering video anomaly detection as a weakly supervised problem. During the training stage, samples with video-level label annotations of normal or abnormal are fed into network. Sultani *et al.* [6] formulated the weakly supervised video anomaly detection (WS-VAD) as a Multiple Instance Learning (MIL) task. Each video is treated as a bag (anomalous: positive normal: negative) and divided into a fixed number of segments as instances. They use Multiple Layer Perceptron (MLP) to generate the anomaly scores for instances and propose a ranking loss to widen the distance between the two highest-scoring instances in the positive and negative bags. After their work was published, many VS-VAD methods like [1][7][8][9] are mainly based on MIL framework.

Zhang *et al.* [7] further introduced inner-bag score gap regularization; AR-Net [8] achieved a better performance as their work introduced a k-max selection method replacing the max selection in previous work [6][7] under the MIL learning

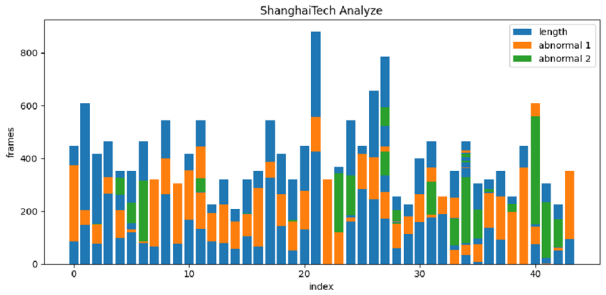


Fig. 2. The duration of abnormal events in ShanghaiTech [2].

framework. Moreover, they introduced dynamic MIL loss (as seen L_{DMIL} in Fig.1.) to maximize inter-class distances. To reinforce the presentation of temporal features of videos, in RTFM [1], MTN was proposed to capture the local dependencies and the global dependencies in temporal dimension between instances.

However, most existing methods did not fully exploit a significant prior knowledge of video anomalies, which is the fact that abnormal events tend to last for a duration in videos. Although the instance represents a clip of consecutive frames in the video, normally it is not the minimum temporal unit for an anomaly event as shown in Fig. 2. Intuitively speaking, introducing the temporal continuity of multiple neighbor instances would be helpful for identifying anomalies. But it also brought another challenge, i.e. it is hard to determine the appropriate duration for the anomaly event while the length of videos is changing at the same time.

To address the aforementioned issues, a Multi-scale Continuity-aware Refinement Network (MCR) is proposed in this work to enforce the WS-VAD tasks pay more attention to the continuity of anomalies. In our work, a new module named Multi-scale Continuity module is applied to extract the continuity of instances in different temporal scales, as shown in Fig.1. Moreover, to better integrate information of different scales, a Multi-scale Attention Module is designed. In this module, specific weighting factors were produced to weight and select the appropriate temporal scales for complex video scenes.

2. METHODOLOGY

Given a video training set of video-labelled samples $V = \{(v_i, y_i)\}_{i=1}^{N}$, where N is the number of videos in training stage, $y_i \in \{0,1\}$ denotes the video-level label annotation indicating whether v_i is a normal video ($y_i = 0$). A single video v_i would be divided into k_i clips. To achieve the goal of discriminating between normal clips and abnormal clips, the anomaly score vector of clips needed to be calculated and could be represented as:

$$S_i = \{s_{i,j}\}_{j=1}^{k_i}, \tag{1}$$

where S_i is the anomaly score vector of v_i and $s_{i,j}$ is the j-th clips anomaly score in S_i.

The overall framework of our work is presented in Fig. 3. A new multi-scale continuity module is proposed to generate window scores which contain the continuous information of instances on different scales, while another multi-scale attention module is utilized to pick the proper scale for anomaly score evaluation.

2.1. Conventional procedures

The key components of conventional WS-VAD include a feature extractor and an anomaly score generator, as shown in the left part of Fig. 3. The input clip for the feature extractor usually consists of 16 consecutive frames, while Inflated 3D(I3D) [11] pretrained on the Kinetics [11] is a widely used one in WS-VAD. The temporal-spatial features F_i could be represented as:

$$F_i = \{f_{i,j}\}_{j=1}^{k_i}, \tag{2}$$

where $f_{i,j}$ is the temporal-spatial features of j-th clips in v_i, $F_i \in \mathbb{R}^{k_i \times D}$, D is the dimension of features.

The feature set F_i is then fed into one Fully Connected (FC) Layer to obtain an enhanced representation $\hat{F}_i$ for anomaly score generation. It is worth noting that this layer has the identical output dimension as the input dimension D. ReLU is adopted as the activation function after the FC Layer. To avoid overfitting, we also adopt Dropout [15] in the training stage. Then $\hat{F}_i$ could be viewed as:

$$\hat{F}_i = \mathcal{D}(\max(0, W_1 F_i + b_1)), \tag{3}$$

where $\mathcal{D}(\cdot)$ denotes Dropout, W_1 and b_1 are parameters of the FC Layer.

After that, $\hat{F}_i$ is sent into a regression layer to generate the basic anomaly score $\dot{S}_i$ as:

$$\dot{S}_i = \frac{1}{1 + \exp(W_r \hat{F}_i + b_r)}, \tag{4}$$

where W_r and b_r are the parameters of the regression layer.

2.2. Multi-scale Continuity-aware Refinement Network

The basic anomaly score $\dot{S}_i$ was used directly to locate anomaly clips in previous work [8][9]. However, this score lacks the continuity information of the video since it is generated individually without considering the context of its corresponding instances. As the continuity is one of the most important characteristics of anomalous events, the proposed MCR network aims to refine the basic anomaly scores $\dot{S}_i$ by introducing the temporal continuity of instances. Thus, the multi-scale continuity module is utilized to extract the anomalous continuity by moving average on various temporal scales. Moreover, multi-scale attention is applied to

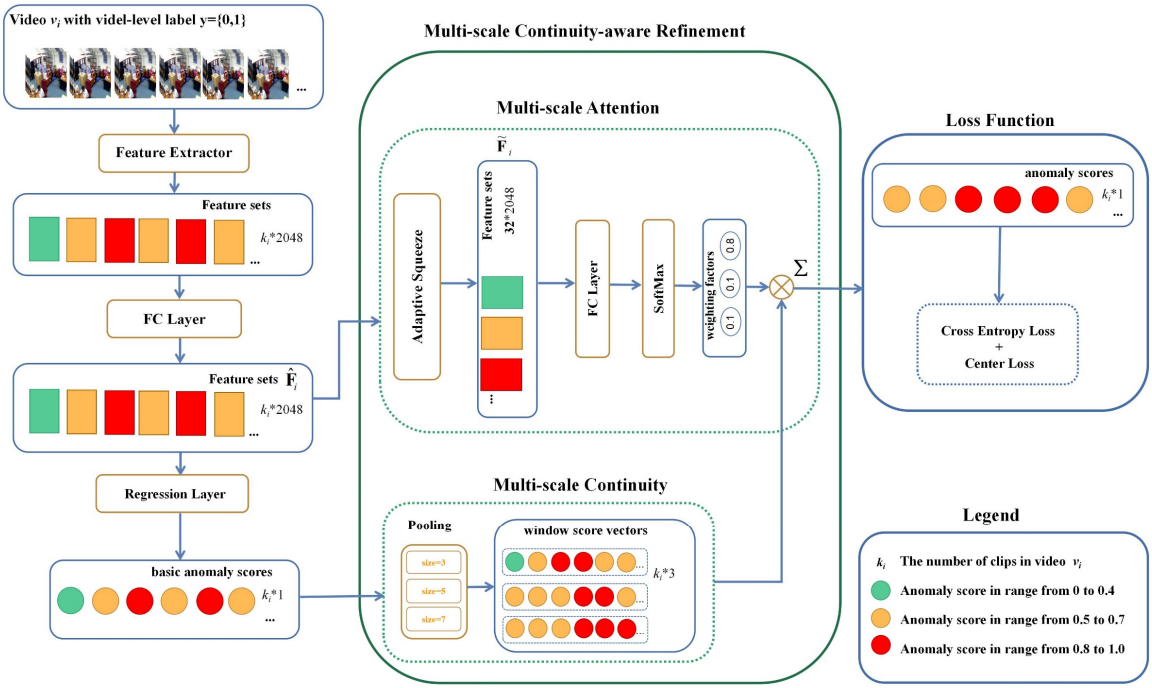


Fig. 3. The overall architecture of the proposed MCR network. The main contribution of this work is marked in the green block. The basic anomaly scores generated by conventional methods are refined according to the rule of multi-scale continuity for final anomaly detection.

better integrate information of different scales in order to enhance the robust of the proposed MCR network.

2.2.1. Multi-scale Continuity

To introduce the temporal continuity of anomalous events based on the basic anomaly score $\dot{S}_i$, the moving average is adopted to generate the window score vectors on different scales. As shown in Fig. 3, the pooling operator is used to perform multi-scale moving average operator on the j-th clips to obtain a new smoothed score $\tilde{s}_{i,j}^{(w)}$,

$$\tilde{s}_{i,j}^{(w)} = \frac{1}{w}\sum_{n=1}^{w} \dot{s}_{i,j+n-1}, \tag{5}$$

where w is the window size of moving average. Compared with RTFM [1], moving average could capture local dependencies of instances without introducing any additional model parameters.

It is noting that the number of clips k_i is inconsistent in each video. Moreover, the duration of different types of anomalous events also varies greatly. Therefore, it is obvious that a fixed temporal scale is not suitable. In the proposed MCR network, a candidate group of score vectors are generated using different window sizes, while the most suitable ones will be picked in the next section. Such set of smoothed scores with N different scales $(w_1, \dots, w_n, \dots, w_N)$ are defined as $\tilde{S}_i = \left\{\tilde{s}_{i,j}^{(w_1)}, \dots, \tilde{s}_{i,j}^{(w_n)}, \dots, \tilde{s}_{i,j}^{(w_N)}\right\}$.

2.2.2. Multi-scale Attention

To select proper scale from the candidate score set $\tilde{S}_i$, an attention-like module is introduced to assign appropriate weight on different scales based on the enhanced feature $\hat{F}_i$ of each video.

Since the number of clips k_i varies from video to video, and adaptive squeeze step is adopted to merge the features of consecutive clips as shown in Fig. 3. After that, the feature $\hat{F}_i \in \mathbb{R}^{k_i \times D}$ can be transferred to $\tilde{F}_i \in \mathbb{R}^{K \times D}$, while K is a hyper-parameter set as 32 in our experiment. Another FC Layer accepts $\tilde{F}_i$ as input, followed by a SoftMax layer as the activation function. Through this procedure, the video-level weighting factors for each scale can be obtained, which indicates the importance of the corresponding smoothed score $\tilde{s}_{i,j}^{(w_n)}$. Then, this weighting factors vector P_i can be formalized as follows,

$$P_i = \left\{p_i^{(n)}\right\}_{n=1}^{N} = \frac{e^{W_2 \hat{F}_i + b_2}}{\sum_j e^{W_2 \hat{F}_j + b_2}}, \tag{6}$$

where $p_i \in [0,1]$. $W_2 \in \mathbb{R}^{D \times N}$, $b_2 \in \mathbb{R}^{1 \times N}$, N indicates the number of weight factors which is the same as the number of scales. It is set as 3 in our experiment. By multiplying the

Table 1. Frame-level AUC of the proposed method against existing methods in ShanghaiTech [2].

Method	Feature extractor	AUC (%)
Sultani *et al*. [6]	C3D	86.30
Zhang *et al.* [7]	C3D	82.50
Zhong *et al.* [12]	C3D	76.44
Zhong *et al.* [12]	TSN^{RGB}	84.44
SRF [16]	C3D	84.17
ClAWS Net [18]	$C3D^{RGB}$	89.67
AR-Net [8] *	$I3D^{RGB}$	88.19
AR-Net [8] *	$I3D^{Conc}$	92.60
MCR	$I3D^{RGB}$	90.10
MCR	$I3D^{Conc}$	**94.92**

* indicate we re-implement the model in our experiment.

smoothed multi-scale anomaly scores $\tilde{S}_i$ with the weight P_i, the final anomaly score $s_{i,j}$ of the *j*-th clip in *i*-th video could be represented as follows,

$$s_{i,j} = \frac{1}{N}\sum_{n=1}^{N} p_i^{(n)} \times \tilde{s}_{i,j}^{(w_n)}, \tag{7}$$

where $p_i^{(n)}$ is the *n*-th constant number in weighting factors P_i, $\tilde{s}_{i,j}^{(w_n)}$ is the score vector of *n*-th scale in the list of score $\tilde{S}_i$. For the sake of simplicity, $S_i = \{s_{i,j}\}_{j=1}^{k_i}$ denotes the score vector of *i*-th video.

2.3. Optimization

Dynamic multiple instances of learning (DMIL) loss and center loss are utilized in the proposed MCR network, which performs well in previous works [8]. With the proposed multi-scale continuity-aware anomaly score $s_{i,j}$, the DMIL loss is capable to enlarge the inter-class distance between anomalous and normal instances, which is represented as follows,

$$\begin{aligned} L_{DMIL} = \frac{1}{k_i}\sum_{s_{i,j}\in S_i}[-y_i \log(s_{i,j}) \\ +(1-y_i)\log(1-s_{i,j})], \end{aligned} \tag{8}$$

where y_i is the video-level label of v_i.

Respectively, the center loss applied here can effectively reduce the intra-class distance, which is formulated as,

$$L_c = \begin{cases} \frac{1}{k_i}\sum_{j=1}^{k_i} \| s_{i,j} - c_i \|, & if\ y_i = 0 \\ 0, & otherwise \end{cases} \tag{9}$$

Table 2. Frame-level AUC of the proposed method against existing methods on UCF-Crime [6] dataset.

Method	Feature extractor	AUC (%)
Sultani *et al*. [6]	C3D	75.41
Zhang *et al.* [7]	C3D	78.70
Zhong *et al.* [12]	C3D	81.08
Zhong *et al.* [12]	TSN^{RGB}	**82.12**
SRF [16]	C3D	79.54
AR-NET [8] *	$I3D^{RGB}$	80.2
MCR	$I3D^{RGB}$	81.0

* indicate we re-implement the model in our experiment.

$$c_i = \frac{1}{k_i}\sum_{j=1}^{k_i} s_{i,j}, \tag{10}$$

where c_i is the center of anomaly score vector s_i of the *j*-th. Then, the total loss function can be represented as follows,

$$L = L_{DMIL} + \lambda L_c. \tag{11}$$

According to the previous frameworks of WS-VAD such as AR-Net [8], λ is set as 20 to keep the balance between two losses in the followed experiments.

3. EXPERIMENTS

3.1. Datasets and metrics

The experiments are conducted on two challenging benchmarks, namely UCF-Crime [6] and ShanghaiTech [2], to evaluate the performance of our proposed MCR network.

UCF-Crime [6] is a large-scale dataset of surveillance videos. It includes 13 types of anomalous behaviors with 1900 long untrimmed videos. 1610 videos with video-level annotation in training stage are available and the other 290 videos with frame-level annotations are test videos.

ShanghaiTech [2] is a medium-scale dataset from campus video surveillance. We followed the procedure as Zhong *et al.* [13], the training splits include 175 normal videos and 63 anomalous while the test splits have 155 normal videos and 44 anomalous videos respectively.

Following the protocol of previous works [16][17][18], the area under the curve (AUC) of the frame-level receiver and operating characteristics (ROC) are used as the evaluation metric for both datasets. Higher AUC indicates better performance. To evaluate the robust of our proposed network, the false alarm rate (FAR) on normal videos is taken

Table 3. Ablation experiment with different scales (pooling sizes in the network).

Scales	AUC (%)	FAR (%)
-	92.60	0.27
3	94.20	0.24
5	94.17	0.29
7	94.15	0.35
9	94.17	0.44
3、5	94.49	0.30
5、7	94.17	0.40
3、5、7	**94.92**	0.28

as another important metric. Lower FAR on normal videos indicates stronger robustness of an anomaly detection method.

3.2. Implementation details

Our proposed MCR network is trained in an end-to-end paradigm using the Adam optimizer [13] with a weight decay of 0.0005 and a batch size of 60, including 30 normal videos and 30 anomalous videos. The learning rate is set to 0.0001 for both ShanghaiTech [2] and UCF-Crime [6]. Following AR-Net [8], we use $I3D^{Conv}$ as our feature presentation, which is combined with RGB frames and Optical-Flow frames generated based on TV-L1 algorithm [14]. We adopted Dropout [15] after the first FC Layer, with a rate of 0.7. And we fixed different clip lengths of videos to 32 by adaptive pooling in the weighting module. In our experiments, the number of weighing factors α is set to be 3, which means we also apply three differential kernel sizes to pool the basic anomaly score. We use 1D average pooling operator to implement moving average, and the kernel sizes of pooling are 3, 5, 7, with padding 1, 2, 3 respectively.

3.3. Comparisons with other methods

The performance of our proposed method on the ShanghaiTech [2] is recorded in Table 1. Different Feature extractors [11][19][20] are adopted in these frameworks. Specially, by adopting the same optimization strategy as AR-Net [8], our method achieved noticeable improvement in terms of AUC metric. The proposed MCR surpass AR-Net on AUC from 92.60% to 94.92% with a relative gain of 2.32%.

Our model is also evaluated on UCF-Crime [6], as shown in Table 2. $I3D^{RGB}$ is used as the features and the same ten-crop augmentation strategy is deployed as AR-Net [8]. Remarkably, our MCR network still obtain a better result than AR-Net [8] on frame-level AUC by 1%, which is also competitive compared to other state-of-the art methods.

3.4. Qualitative results and analysis

The visualized results of both normal and anomaly score plots generated by our method and AR-Net [8] are compared in Fig. 4. It can be seen that the results of our MCR tend to be more continuous, which is just in line with the characteristics of anomalous events. In the normal video (Fig. 4(a)), the incorrect high anomaly score by AR-Net [8] around the 250-th frame, can be effectively pulled down by neighboring frames in our model. Respectively, the score plots of anomaly events in (b) and (c) given by MCR are closer to the ground truth.

3.5. Ablation study

In order to verify the effectiveness of multi-scale strategy, AR-Net [8] is used to generate the basic anomaly score as baseline, then a set of experiments is conducted using different window size.

As shown in Table 4, with the increase in pooling size, the frame-level AUC changes subtly, but the false positive rate always increases. Considering the robust of our proposed network, we would prefer the smaller three (size on 3, 5, 7).

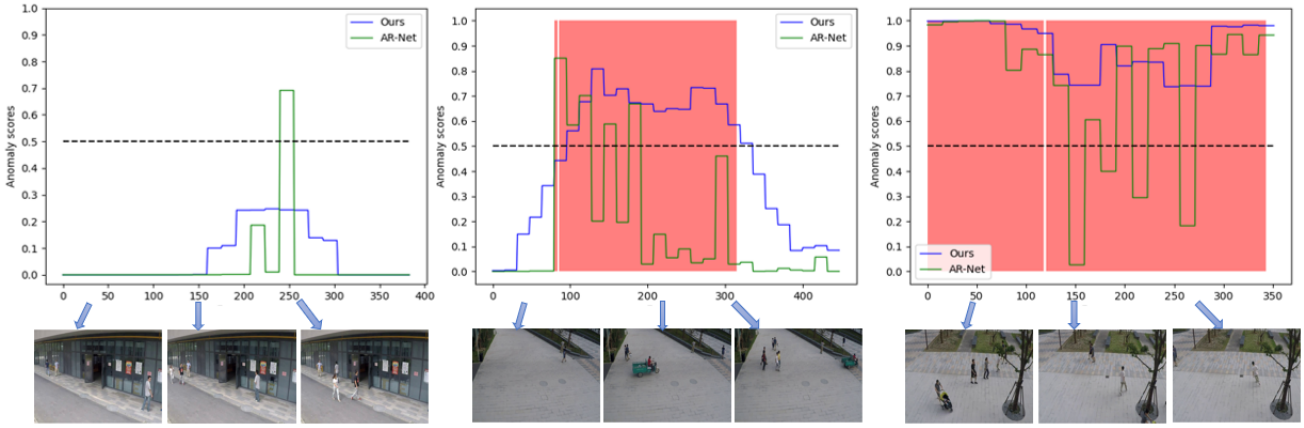
The ablation study shows that the proposed multi-scale scheme is effective to raise the frame-level AUC to 94.12%. Moreover, it boosts the performance to a frame-level AUC 94.92% after we fuse the multi-scale continuous information by weighting factors, with the improvement about 2.32% towards baseline.

4. CONCLUSION

In this paper, we presented a Multi-scale Continuity-aware Refinement Network (MCR) for video anomaly detection. Our network follows weakly supervised video anomaly detection paradigm, with only video-level label available in the training stage. Different from previous work based on MIL-learning, we fully exploit the temporal continuity characteristic of anomaly events. We extract multi-scale continuity information on neighbor clips and weight it by Multi-scale Attention Module. Experiments on challenging datasets demonstrate the effectiveness of our method.

5. ACKNOWLEDGEMENT

This work was supported by the Zhejiang Provincial Natural Science Foundation of China (No. LY20F030005) and National Natural Science Foundation of China (No. 61603202).



(a) 06_009 (b) 01_0053 (c) 04_0012

Fig. 4. Visualized examples of the proposed MCR on ShanghaiTech [2]. (a) does not contain any anomalous events. Red regions in (b) and (c) denotes the anomalous ground truth.

6. REFERENCES

[1] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," In *arXiv preprint*, 2021.

[2] W. Luo, W. Liu and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," In *ICCV*, 2017, pp. 341-349.

[3] M. Sabokrou, M. Khalooei, M. Fathy and E. Adeli, "Adversarially learned one-class classifier for novelty detection," In *CVPR*, 2018, pp. 3379-3388.

[4] T. N. Nguyen and J. Meunier, "Anomaly detection in video sequence with appearance-motion correspondence," In *ICCV*, 2019, pp. 1273-1283.

[5] H. Park, J. Noh and B. Ham, "Learning memory-guided normality for anomaly detection," In *CVPR*, 2020, pp. 14372-14381.

[6] W. Sultani, C. Chen and M. Shah, "Real-world anomaly detection in surveillance videos," In *CVPR*, 2018, pp. 6479-6488.

[7] J. Zhang, L. Qing and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," In *ICIP*, 2019, pp. 4030-4034.

[8] B. Wan, Y. Fang, X. Xia and J. Mei, "Weakly supervised video anomaly detection via center-guided discriminative learning," In *ICME*, 2020, pp. 1-6.

[9] S. Yu, C. Wang, Q. Mao, Y. Li and J. Wu, "Cross-Epoch Learning for Weakly Supervised Anomaly Detection in Surveillance Videos," In *SPL*, 2021, vol. 28, pp. 2137-2141.

[10] K. Dohun, K. Heegwang, M. Yeongheon and J. Paik, "Real-Time Surveillance System for Analyzing Abnormal Behavior of Pedestrians," In *Applied Sciences*, 2021, pp. 6153.

[11] C. Joao and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," In *CVPR*, 2017, pp. 6299-6308.

[12] J. 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 *CVPR*, 2019, pp. 1237-1246.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," In *arXiv preprint*, 2014.

[14] F. Steinbrücker, T. Pock and D. Cremers, "Large displacement optical flow computation withoutwarping," In *ICCV*, 2009, pp. 1609-1614.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," In *JMLR*, 2014, pp. 1929-1958.

[16] M. Z. Zaheer, A. Mahmood, H. Shin and S. I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," In *IEEE Signal Processing Letters*, 2020, pp. 1705-1709.

[17] J. C. Feng, F. T. Hong and W. S. Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," In *CVPR*, 2021, pp. 14009-14018.

[18] M. Z. Zaheer, A. Mahmood, M. Astrid and S. I. Lee, "CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," In *ECCV*, 2020, pp. 358-376.

[19] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang and L. V. Gool, "Temporal segment networks for action recognition in videos," In *PAML*, 2018, pp. 2740-2755.

[20] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," In *ICCV*, 2015, pp. 4489-4497.