WEAKLY SUPERVISED VIDEO ANOMALY DETECTION VIA CENTER-GUIDED DISCRIMINATIVE LEARNING

Boyang Wan, Yuming Fang, Xue Xia, Jiajie Mei

School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China

ABSTRACT

Anomaly detection in surveillance videos is a challenging task due to the diversity of anomalous video content and duration. In this paper, we consider video anomaly detection as a regression problem with respect to anomaly scores of video clips under weak supervision. Hence, we propose an anomaly detection framework, called Anomaly Regression Net (AR-Net), which only requires video-level labels in training stage. Further, to learn discriminative features for anomaly detection, we design a dynamic multiple-instance learning loss and a center loss for the proposed AR-Net. The former is used to enlarge the inter-class distance between anomalous and normal instances, while the latter is proposed to reduce the intra-class distance of normal instances. Comprehensive experiments are performed on a challenging benchmark: *ShanghaiTech*. Our method yields a new state-of-the-art result for video anomaly detection on *ShanghaiTech* dataset.

Index Terms— Anomaly detection, weak supervision, multiple-instance learning, center loss

1. INTRODUCTION

Video anomaly detection is an important yet challenging task in computer vision, and it is widely used in crime warning, intelligent video surveillance and evidence collection. According to the study [1], there are two kinds of paradigms: unary classification and binary classification, for weakly-supervised video anomaly detection. Anomalies are usually defined as the video content patterns that are different from usual patterns in previous works [2] [3] [4] [5] [6]. Based on this definition, the unary classification paradigm-based methods only model usual patterns with normal training samples. However, it is impossible to collect all kinds of normal samples in a training set. Consequently, normal videos being different from training ones may tend to be false alarmed under this paradigm.

To address this issue, the binary classification paradigm was introduced, in which training data contains both anomalous and normal videos. Following the binary classification
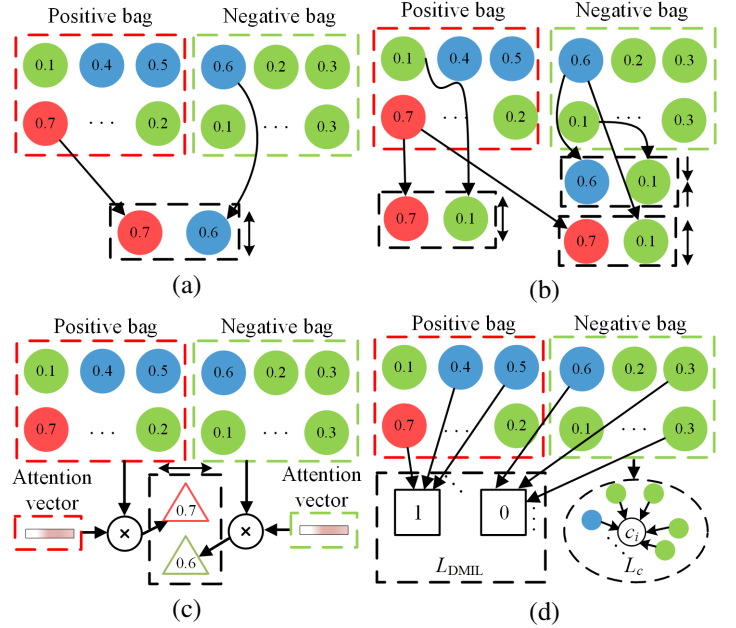
Fig. 1. Comparison of loss functions for MIL-based anomaly detection methods. Float numbers in colored circles stand for anomaly scores of segments or clips, while those in colored triangles represent anomaly scores of videos. Binary numbers in black squares are video-level labels. (a) MIL ranking loss in [7]. (b) Complementary inner bag loss in [9]. (c) Temporal ranking loss in [8]. (d) Ours.

paradigm, some studies on anomaly detection [1] [7] [8] [9] have been published. In [1], video anomaly detection was formulated as a fully-supervised learning task under noise labels. As a correction, a Graph Convolutional Network (GCN) was proposed to train an action classifier. The GCN and the action classifier were optimized alternately.

In this paper, we formulate video anomaly detection as a weakly-supervised learning problem following binary classification paradigm, where only video-level labels are involved in training stage. Recently, Multiple-Instance Learning (MIL) has become a major technique in several computer vision tasks including weakly-supervised temporal activity localization and classification [10] [11] and weakly-supervised object detection [12]. There are also some MIL-based video

anomaly detection studies [7] [8] [9]. Each training video is treated as a bag, and clips of videos are regarded as instances in these methods. An anomalous video is treated as a positive bag, and a normal one is presented as a negative bag. Sultani *et al.* [7] proposed a deep MIL ranking model with features extracted by C3D network [13] as input. The deep MIL ranking loss was proposed for separating the anomaly scores of anomalous and normal instances. Zhang *et al.* [9] proposed a complementary inner bag loss to reduce intra-class distances and enlarge inter-class distances of instances simultaneously. The highest and lowest anomaly scores of anomalous and normal videos were required. Zhu and Newsam [8] proposed a temporal ranking loss calculated according to anomaly scores. The anomaly score of a video is the weighted sum of a video anomaly score vector and an attention vector. However, as shown in Fig 1, these methods adopted pair-wisely calculated losses, based on which the detection ability of models partly depend on batch size. In other words, the detection performance is partly limited by graphics memory. In this work, we look into a method that maximizes the inter-class distances and minimizes intra-class distances without instances from pair videos.

We propose a framework, termed Anomaly Regression Network (AR-Net), and two novel losses to learn discriminative features under video-level weak supervision. As illustrated in Fig 1-(d), the dynamic multiple-instance learning loss (L_{DMIL}) is proposed to make features more separable. L_{DMIL} is acquired by calculating the cross entropy between the anomaly scores of video clips and their corresponding video labels. The center loss is designed as the distances between anomaly scores of video clips and their corresponding average anomaly score c_i in each training normal video. By minimizing the two losses, a discriminative feature representation can be obtained for video anomaly detection.

Comprehensive experiments are conducted on a benchmark: ShanghaiTech [14]. Our approach yields a new state-of-the-art result and obtains an absolute gain of 4.94% in terms of Area Under the Curve (AUC) on ShanghaiTech dataset.

2. PROPOSED METHOD

In this section, we first define the notations and problem statement. Then we describe the proposed feature extraction network. Finally, we present our AR-Net followed by a detailed description of the proposed losses.

Problem Statement: A training set consisting of n videos is denoted by $\chi = \{x_i\}_{i=1}^n$ in an anomaly detection dataset. The temporal duration of the dataset is defined as $T = \{t_i\}_{i=1}^n$, where t_i is the clip number of the i-th video. The video anomaly label set is denoted as $Y = \{y_i\}_{i=1}^n$, where $y_i = \{0, 1\}$. In the testing stage, the predicted anomaly score vector of a video x is denoted as $s = \{s^j\}_{j=1}^t$, where $s^j \in [0, 1]$, and s^j is anomaly score of the j-th video clip.

2.1. Feature Extraction

To make use of both the appearance and motion information of the videos, Inflated 3D (I3D) [15], pretrained on the Kinetics [15] dataset, is used as the feature extraction network. An input video is divided into non-overlapped clips, each of which contains 16 consecutive frames. The RGB and Optical-Flow versions of I3D are denoted by I3D$^{\text{RGB}}$ and I3D$^{\text{Optical-Flow}}$ respectively. The former takes RGB frames as input while the latter takes Optical-Flow frames. We concatenate features from penultimate layer of the I3D$^{\text{RGB}}$ and I3D$^{\text{Optical-Flow}}$ as our final feature representation of video clips.

The feature matrix X_i shown in Fig 2 is composed of features from the training video x_i. The dimension of X_i is $F \times t_i$, where F is the dimension of clip features. X_i rather than raw video clips is fed to our AR-Net.

2.2. Anomaly Regression Network

The architecture of AR-Net is shown in Fig 2. The Fully Connected Layer (FC-Layer) and Anomaly Regression Layer (AR-Layer) in AR-Net require only video-level labels for video anomaly detection. We adopt ReLU [17] as the activation function of FC-Layer. To avoid overfitting, Dropout [18] is introduced to FC-Layer and can be formalized as follows:

$$X_i^{\text{FC}} = D(max(0, W_{\text{FC}}X_i + b_{\text{FC}})) \qquad (1)$$

where $D(\cdot)$ denotes Dropout, $W_{\text{FC}} \in \mathbb{R}^{F \times F}$ and $b_{\text{FC}} \in \mathbb{R}^{F \times 1}$ are learnable parameters to be optimized from the training data, and $X_i^{\text{FC}} \in \mathbb{R}^{F \times t_i}$ is the i-th video feature output from output features of the FC-Layer.

We establish a mapping function between the representations X_i^{FC} and anomaly score vectors s_i by AR-Layer, which is a fully connected layer. The AR-Layer can be represented as follows:

$$s_i = \frac{1}{1 + exp(W_{\text{AR}}X_i^{\text{FC}} + b_{\text{AR}})} \qquad (2)$$

where $W_{\text{AR}} \in \mathbb{R}^{1 \times F}$, $b_{\text{AR}} \in \mathbb{R}^1$ are learnable parameters and $s_i \in \mathbb{R}^{1 \times t_i}$. The anomaly score vectors s_i represent the probabilities that instances are classified as anomalies.

2.3. Dynamic Multiple-Instance Learning Loss

As discussed in Section 1, the video anomaly detection is treated as a MIL task in this paper. In MIL, a positive bag contains at least one positive instance and a negative bag contains no positive instances, *i.e.*, an abnormal video contains at least one anomalous event and a normal video contains no anomalous events. To enlarge the inter-class distance between anomalous and normal instances under a weak supervision, inspired by the k-max MIL loss in [10] [11], we propose a Dynamic Multiple-Instance learning (DMIL) loss that takes the diversity of video duration into consideration.
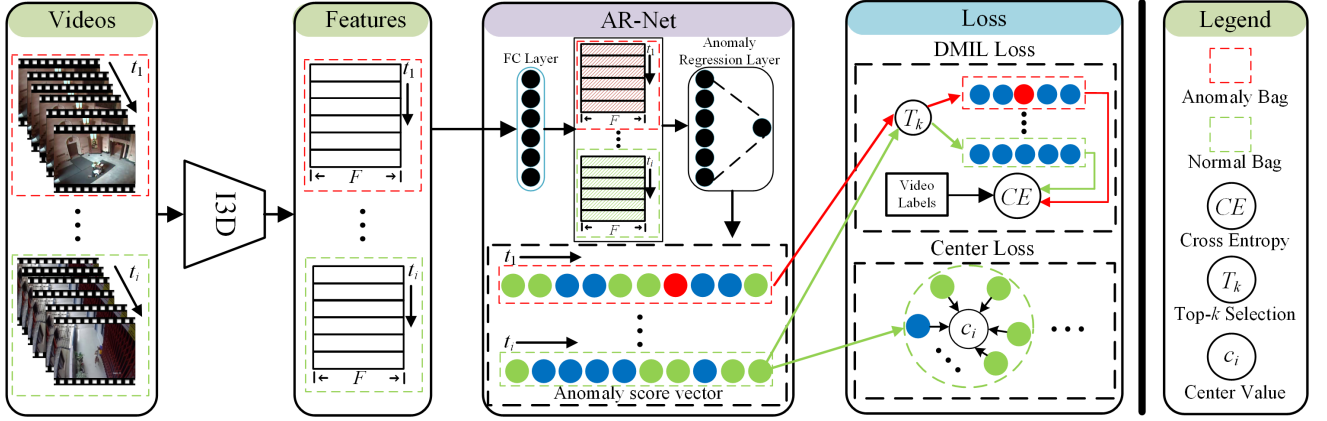
Fig. 2. Our model (AR-Net) with two proposed loss terms (DMIL, center), feature extractor and legend.

Different from the max-selection method involved in MIL-based loss function used in [7] [8] [9], we introduce k-max selection method, which is used in [10] [11], to obtain the k-max anomaly scores. The k is determined based on the number of clips in a video. Specifically,

$$k_i = \lceil \frac{t_i}{\alpha} \rceil \tag{3}$$

where α is a hyperparameter. Thus, the k-max anomaly scores of the i-th video can be represented as,

$$\begin{cases} \boldsymbol{p}_i = sort(\boldsymbol{s}_i) \\ \boldsymbol{S}_i = \{p_i^j \mid j = 1, 2, ..., k_i\} \end{cases} \tag{4}$$

where $\boldsymbol{s}_i$ is anomaly score vector of the i-th video, $sort(\cdot)$ is a descending sort operator and $\boldsymbol{p}_i$ is the sorted $\boldsymbol{s}_i$. Thus, $\boldsymbol{S}_i$ consists of top-k_i elements in $\boldsymbol{s}_i$. The DMIL loss can then be represented as follows:

$$\begin{aligned} L_{\text{DMIL}} = \frac{1}{k_i} \sum_{s_i^j \in \boldsymbol{S}_i} [-y_i log(s_i^j) \\ + (1 - y_i) log(1 - s_i^j)] \end{aligned} \tag{5}$$

where $y_i = \{0, 1\}$ is the video anomaly label. Furthermore, instead of calculating the cross-entropy between the average of selected k scores and the video label in [10] [11], we calculate the cross-entropy between each of the selected k scores and the video label as the instance loss respectively. Noise labels will affect anomaly scores of the sample features from which an average anomaly score is calculated. While our DMIL loss focuses on individual anomaly scores rather than an average one. Thus, this loss keeps errors brought by noise labels from propagating.

2.4. Center loss for Anomaly Scores Regression

The objective of the DMIL loss is to enlarge the inter-class distance of instances. However, both the max and k-max selection method inevitably produce wrong label assignment, since the anomaly scores of normal clips and abnormal clips in the abnormal video are similar in early training stage. As a result, the intra-class distance of normal instances is unfortunately enlarged by the DMIL loss, and this will reduce detection accuracy in testing stage.

Inspired by the center loss in [19], we propose a novel center loss for anomaly score regression to address the above-mentioned issue. In [19], the center loss learns the feature center of each class and penalizes the distance between the feature representations and their corresponding class centers. In our case, the center loss proposed for anomaly score regression gathers the anomaly scores of normal video clips. Our center loss for anomaly score regression can be represented as,

$$L_c = \begin{cases} \frac{1}{t_i} \sum_{j=1}^{t_i} \left\| s_i^j - c_i \right\|_2^2, & \text{if } y_i = 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$c_i = \frac{1}{t_i} \sum_{j=1}^{t_i} s_i^j \tag{7}$$

where c_i is the center of anomaly score vector $\boldsymbol{s}_i$ of the i-th video.

2.5. Optimization

The total loss function of the AR-Net can be represented as follows:

$$L = L_{\text{DMIL}} + \lambda L_c \tag{8}$$

3

To achieve a balance between the two losses in training stage, we empirically set $\lambda = 20$.

3. EXPERIMENTS

3.1. Experiments Setup

Datasets: The proposed AR-Net is evaluated on a challenging dataset containing untrimmed videos with variable scenes, contents and durations. ShanghaiTech [14] is a dataset that contains 437 videos with 130 anomalies on 13 scenes. However, this dataset is proposed for unary-classification, so all the training videos are normal [14]. To make binary-classification available, we adopt the split version proposed in [1]. Specifically, there are 238 training videos and 199 testing videos.

Evaluation Metric: Similar to previous works [1] [14] [7], we use an Area Under of Curve (AUC) of the frame-level Receiver Operating Characteristics (ROC) and False Alarm Rate (FAR) with threshold 0.5 as the evaluation metrics. In the video anomaly detection task, the higher AUC demonstrates, the better the model performs, and lower FAR on a normal video implies stronger robustness of an anomaly detection method.

Implementation Details: We combine I3D^{RGB} and I3D$^{Optical-Flow}$ as our feature extractor, denoted as I3D^{Conc}. The feature of I3D^{Conc} is the concatenation of features from I3D^{RGB} and I3D$^{Optical-Flow}$. Besides, our feature extractor is not fine-tuned. The Optical-Flow frames of each clip are generated based on TV-L1 algorithm [16]. Empirically, we set $\alpha = 4$ for ShanghaiTech dataset. The weights of the AR-Net are initialized by Xavier method [20], and the Dropout probability for FC-Layer is 0.7. We adopt the Adam optimizer [21] with a batch size of 60, in which 30 normal videos and 30 abnormal ones are randomly selected from the training set. The learning rate is always 10^{-4} in our experiments.

Table 1. AUC and FAR of the proposed method against 3 existing methods. The §, † and ‡ indicate the anomaly detection model based on C3D, TSN^{RGB} and TSN$^{Optical-Flow}$ in [1], respectively.

Methods	ShanghaiTech	
	AUC(%)	FAR(%)
Sultani *et al.* [7]	86.30	0.15
Zhang *et al.* [9]	82.50	0.10
Zhong *et al.*§ [1]	76.44	—
Zhong *et al.*† [1]	84.44	—
Zhong *et al.*‡ [1]	84.13	—
AR-Net	**91.24**	**0.10**

3.2. Comparison Results

Table 1 shows the comparison of our method against existing approaches [7] [1] [9] on the ShanghaiTech.

In order to present a comparison against MIL-based works on ShanghaiTech, we reproduced the method in [9] and adopted the open source code provided by Sultani *et al.* [7] to conduct the anomaly detection. The above two models are obtained by pretrained C3D. As shown in Table 1, [9] achieves a frame-level AUC of 82.50%. Meanwhile, [7] performs a frame-level AUC of 86.30%, outperforming the best existing method [1]. Our method substantially exceeds both Sultani *et al.* [7] and Zhong *et al.*† [1] with a frame-level AUC of 91.24%. Furthermore, our approach is the only one surpassing 90% in terms of AUC on ShanghaiTech.

Table 2. AUC and FAR of different loss functions.

Losses	ShanghaiTech	
	AUC(%)	FAR(%)
Baseline:$L_{k\text{-max MIL}}$	86.50	0.93
Ours:L_{DMIL}	89.10	0.21
Ours:$L_{\text{DMIL}} + L_c$	91.24	0.10

Table 3. AUC and FAR of different feature extractors.

| Feature extractor	ShanghaiTech	
	AUC(%)	FAR(%)
I3D^{RGB}	85.38	0.27
I3D$^{Optical-Flow}$	82.34	0.37
I3D^{Conc}	91.24	0.10

3.3. Ablation Study

The comparison results by using different loss functions in Table 2 illustrate the boost brought by the proposed L_c and L_{DMIL} in our AR-Net. AR-Net that involves k-max selection-based MIL loss ($L_{k\text{-max MIL}}$) is treated as the baseline in our ablation study. It achieves a frame-level AUC of 86.50% on ShanghaiTech. While the proposed DMIL loss-based AR-Net boosts the performance by obtaining a frame-level AUC of 89.10% on ShanghaiTech. Besides, with the help of the proposed L_c, the FAR on ShanghaiTech is reduced to 1/9 of those achieved by baseline.

To demonstrate the performance brought by video appearance and motion information, we compare the anomaly detection results based on different feature extractors. As show in Table 3, AR-Net with I3D^{RGB} achieves a frame-level AUC of 85.38%. And the I3D$^{Optical-Flow}$ based AR-Net achieves a frame-level AUC of 82.34%. The AR-Net with I3D^{Conc} boosts the performance with a frame-level AUC of 91.24%.

3.4. Qualitative Analysis

As shown in Table 1, our method surpasses most of the recent MIL-based works. In ShanghaiTech, sometimes anomaly makes up a tiny part of a whole video. In segmented-based methods [7] [9], although each video is divided into 32 non-overlapped segments, the anomalous events in each segment

4

(a) 01_0177   (b) 01_0015
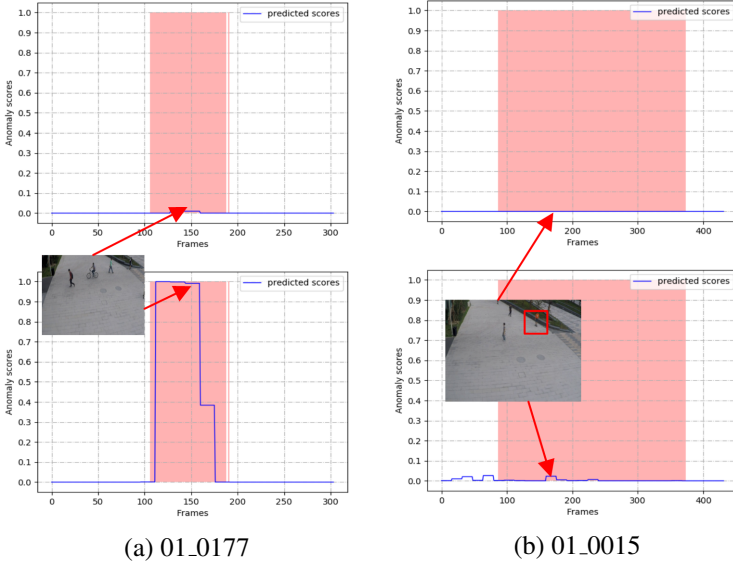
Fig. 3. Comparison visualization of testing results between [7] and ours.

Fig. 4. AUC of different α values.

may still account for little parts. In other words, the features of anomalous frames to be overwhelmed by normal ones in a segment. As a result, segments containing anomaly tend to be regarded as normal patterns, i.e., the anomaly detection model trained by these segments lacks capability to detect short-term anomaly. While in clip-based methods like [10] [11] and ours, videos are divided into clips with fixed number of frames. Anomaly detection models using this strategy avoids anomaly being overwhelmed by normal frames and are able to recognize short-term anomaly. As shown in Fig.3-(a), the model in [7] (illustrated in the 1st row) does not recognize illegal bicycling while ours (in the 2nd row) does.

As shown in the 1st row of Fig.3-(b), both of [7] (illustrated in the 1st row) and our method (in the 2nd row) fail to detect abnormal events in '01_0015'. The reasons are: 1) the anomaly, marked with a red box, only takes up a local part of the video scene, while methods with global input are prone to ignoring local anomalies 2) the anomaly of skateboarding on the sidewalk are not visually distinguishable from normal behaviors, hence the anomalous clips and non-ones are not separable. In conclusion, video anomaly detection in these kinds of scenes is still a big challenge for current models.

In order to gain insight into the hyperparameter α, we perform experiments using the I3D$^{\text{Conc}}$ feature extractor with different values of α, as shown in Fig.4. In fact, the α determines the proportion of noisy-label instances in training stage. Although higher α results in a smaller proportion of noisy-label instances, some anomalous instances will be ignored in training stage. This leads to insufficient diversity of training anomalous instances, which reduces the frame-level AUC of the AR-Net. There is at most 2.44% decrease on frame-level AUC when α is set to 8, 16, 32, or 64. Moreover,
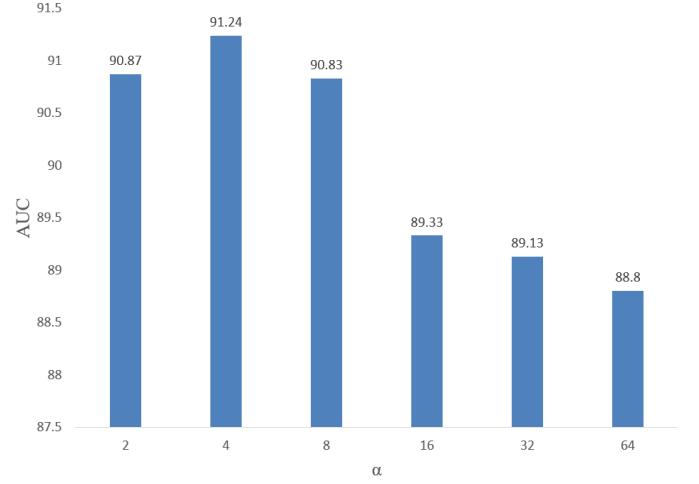
the lower α causes more normal instances being labeled as the anomalous ones in training stage. This also reduces the frame-level AUC of the AR-Net. Fig 4 shows that AR-Net suffers 0.37% decrease on frame-level AUC when α is lower than 4.

4. CONCLUSION

In this paper, we propose a MIL-based anomaly regression network for video anomaly detection. Besides, we design a dynamic loss L_{DMIL} to learn separable features and a center loss L_c to correct the anomaly scores output by our AR-Net. By optimizing the parameters of AR-Net under weak supervision, the dynamic multiple-instance learning loss avoids false alarm caused by interference between clip features. While the center regression loss suppresses label noise by smoothing the distribution of anomaly scores. In addition, the clip-based instance generation strategy benefits to short-term anomaly detection. Experiments on a challenging datasets clearly demonstrate the effectiveness of our approach for video anomaly detection. In the future, we will investigate to model temporal relation between instances to obtain more robustness.

5. REFERENCES

[1] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1237–1246.

[2] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao, "Future frame prediction for anomaly detection–a new baseline," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6536–6545.

5

[3] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao, "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7842–7851.

[4] Cewu Lu, Jianping Shi, and Jiaya Jia, "Abnormal event detection at 150 fps in matlab," in *IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727.

[5] Weixin Luo, Wen Liu, and Shenghua Gao, "Remembering history with convolutional lstm for anomaly detection," in *IEEE International Conference on Multimedia and Expo*, 2017, pp. 439–444.

[6] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis, "Learning temporal regularity in video sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.

[7] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6479–6488.

[8] Yi Zhu and Shawn D. Newsam, "Motion-aware feature for improved video anomaly detection," *CoRR*, vol. abs/1907.10211, 2019.

[9] Jiangong Zhang, Laiyun Qing, and Jun Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *IEEE International Conference on Image Processing*, 2019, pp. 4030–4034.

[10] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury, "W-TALC: Weakly-supervised temporal activity localization and classification," in *European Conference on Computer Vision*, 2018, pp. 563–579.

[11] Sanath Narayan, Hisham Cholakkal, Fahad Shahbaz Khan, and Ling Shao, "3C-Net: Category count and center loss for weakly-supervised action localization," in *IEEE International Conference on Computer Vision*, 2019, pp. 8679–8687.

[12] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye, "C-MIL: Continuation multiple instance learning for weakly supervised object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2199–2208.

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4489–4497.

[14] Weixin Luo, Wen Liu, and Shenghua Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *IEEE International Conference on Computer Vision*, 2017, pp. 341–349.

[15] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[16] Frank Steinbrcker, Thomas Pock, and Daniel Cremers, "Large displacement optical flow computation without-warping," in *IEEE International Conference on Computer Vision*, 2009, pp. 1609–1614.

[17] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.

[18] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[19] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, 2016, pp. 499–515.

[20] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.

[21] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.