LEARNING TASK-SPECIFIC REPRESENTATION FOR VIDEO ANOMALY DETECTION WITH SPATIAL-TEMPORAL ATTENTION

*Yang Liu[1], Jing Liu[1], Xiaoguang Zhu[2], Donglai Wei[1], Xiaohong Huang[3], Liang Song[1]**

[1]Academy for Engineering & Technology, Fudan University, Shanghai, China
[2]SEIEE, Shanghai Jiao Tong University, Shanghai, China
[3]SIME, Shanghai University of Finance and Economics, Shanghai, China

ABSTRACT

The automatic detection of abnormal events in surveillance videos with weak supervision has been formulated as a multiple instance learning task, which aims to localize the clips containing abnormal events temporally with the video-level labels. However, most existing methods rely on the features extracted by the pre-trained action recognition models, which are not discriminative enough for video anomaly detection. In this work, we propose a spatial-temporal attention mechanism to learn inter- and intra-correlations of video clips, and the boosted features are encouraged to be task-specific via the mutual cosine embedding loss. Experimental results on standard benchmarks demonstrate the effectiveness of the spatial-temporal attention, and our method achieves superior performance to the state-of-the-art methods.

***Index Terms*—** Video anomaly detection, multiple instance learning, spatial-temporal attention, deep learning

1. INTRODUCTION

Video anomaly detection (VAD) aims to detect abnormal events such as criminal behaviors in surveillance videos, and it is widely used in intelligent surveillance systems [1]. Traditional manual detection is time-consuming and can lead to missed detection due to visual fatigue. Therefore, the automatic detection of abnormal events is of great practical value. Since abnormal events are rare and diverse, it is almost impossible to collect and label all kinds of anomalies for modeling [2]. Therefore, most of the existing methods formulate VAD as an unsupervised [3, 4, 5, 6] or a weakly supervised task [2, 7, 8, 9, 10]. Unsupervised methods use normal samples to learn a model of 'normality', and the anomaly is detected by measuring its deviation to the learned model. Due to the lack of observation of abnormal events, the unsupervised models may not learn the essential difference between normal and anomaly, leading to false alarms. In contrast, the weakly supervised VAD (ws-VAD) detects anomalies by comparing the normal and abnormal clips with the video-level labels, which can produce more reliable results [8].

The ws-VAD has been formulated as a multiple instance learning (MIL) [11] task, which uses the easy-to-obtain video-level labels to localize the clips where abnormal behaviors occur, avoiding the high cost of clip-level annotations. Sultani *et al.* [2] firstly proposed a MIL ranking model to calculate anomaly scores. The objective is that the maximum score of clips from an abnormal video should be greater than that of a normal video. Zhu *et al.* [7] proposed a temporal network to learn the motion-aware feature of video clips and fed it to an attention-based MIL ranking model. Zhong *et al.* [12] applied the graph convolutional network to clean the label noise and used the cleaned labels to supervise the MIL ranking model. Feng *et al.* [8] proposed a self-training framework to learn task-specific representation with a self-guided attention feature encoder. Most of the existing methods [2, 12] directly use the spatial-temporal features extracted by the pre-trained models, such as convolutional 3D (C3D) [13] and inflated 3D ConvNet (I3D) [14]. However, the pre-trained models are designed for action recognition tasks instead of VAD, so the extracted features are not discriminative enough to distinguish normal and abnormal events. In addition, previous works [2, 7] always treat the video clips cut from the same video as independent instances, ignoring the inter-connections between adjacent video clips.

To obtain the task-specific spatial-temporal features for ws-VAD, we propose spatial-temporal attention (STA) to explore the inter- and intra-correlations between video clips. The STA module can capture the global contextual spatial-temporal correlations through a recurrent crisscross attention [15] operation. Considering the diversity of normal events, we further introduce a mutual cosine loss to obtain more compact representations of normal clips while keeping the representation of abnormal clips away. The main contributions of this work are summarized as follows:

- We propose spatial-temporal attention to obtain task-specific features for ws-VAD. The global spatial-temporal correlations of all video clips can be captured via the easy-to-plugin recurrent attention operations.

*Corresponding author. This work is suppoted by the Shanghai Key Research Laboratory of NSAI.

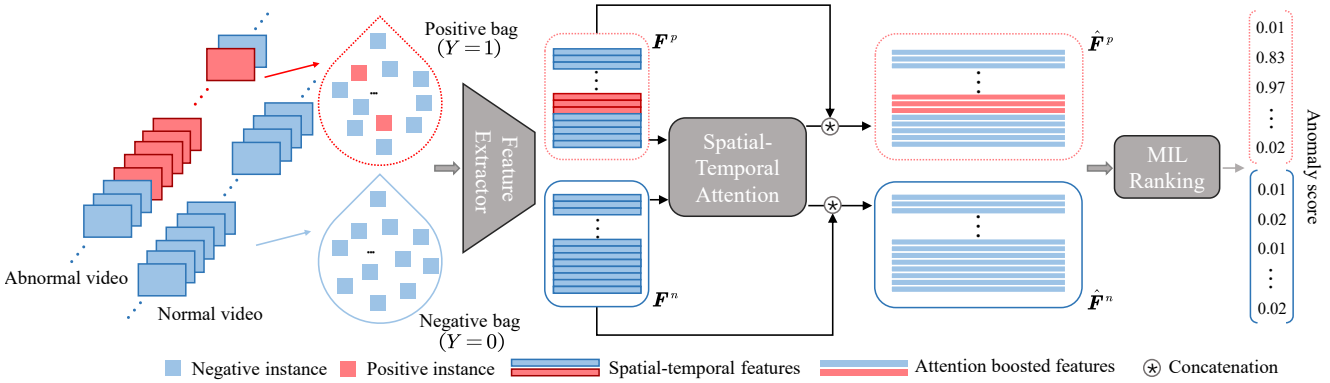

Fig. 1. Architecture of the spatial-temporal attention augmented MIL ranking framework.

- We propose an STA augmented MIL ranking model and introduce a mutual cosine loss to encourage the model to learn the prototypical patterns of normal events.
- Experimental results on three standard benchmarks demonstrate the effectiveness of the STA, and our model outperforms the state-of-the-art methods.

2. METHODS

2.1. Architecture

The architecture of the proposed STA augmented MIL ranking framework is shown in Figure 1, including a feature extractor, a spatial-temporal attention module, and a MIL ranking module. Firstly, each video is cut into N non-overlapping clips, and a clip consists of K consecutive frames. A video containing anomalies is labeled as positive ($Y = 1$) and represented as a positive bag containing N instances, while a video without any anomaly is represented as a negative bag. To avoid additional training overhead, we apply the well-trained C3D [13] or I3D [14] model as the feature extractor, which encodes the instances into feature vectors, denoted by $\boldsymbol{f}$. Considering that the $\boldsymbol{f}$ are interrelated and are not sufficiently discriminatory for ws-VAD, so we feed $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_N\}$ into the STA module to capture the global spatial-temporal correlations through a recurrent criss-cross attention. To further capture the temporal correlations between video clips, we replace the fully connected (FC) network in the MIL ranking model with a bi-direction recurrent neural network (bd-RNN), which takes the boosted features $\{\hat{\boldsymbol{f}}_1, \hat{\boldsymbol{f}}_2, \cdots, \hat{\boldsymbol{f}}_N\}$ as input and outputs the anomaly score of each instance at a range of $[0, 1]$.

2.2. Spatial-temporal attention

Expanding the spatial-temporal features of all clips cut from the same video in space will form a feature map $\boldsymbol{F}$ of size $N \times C$, where N and C denote the number of clips and dimension of the feature vector, respectively. Since the pre-training model is applied to each video clip individually, there is no connection between the individual vectors in $\boldsymbol{F}$. The pixel $\boldsymbol{F}_{(i,j)}$ is considered as the feature of i-th clip in a specific local space-time region. The other pixels should be relevant with $\boldsymbol{F}_{(i,j)}$, and the magnitude of the correlation is positively related to the spatial-temporal distance, which is reflected as the spatial distance on the feature map. Therefore, we introduce crisscross attention to aggregate the contextual information between each local spatial-temporal feature, and the global correlations can be captured by the recurrent cross-attention operation.

Specifically, the details of the spatial-temporal attention are shown in Figure 2. Firstly, we obtain the query map and key map via the 1×1 convolution, denoted by $\boldsymbol{Q}$ and $\boldsymbol{K}$, respectively, and the size is $N \times C \times D$, where D is the number of channels. The vector of the i-th row and j-th column of query map $\boldsymbol{Q}$ is denoted by $\boldsymbol{q}_{(i,j)}$, obviously $1 \leq i \leq N, 1 \leq j \leq C$ and $\boldsymbol{q}_{(i,j)} \in \mathbb{R}^D$, then we obtain criss-cross attention map $\boldsymbol{A}^{i,j}$ by computing the cosine similarity between $\boldsymbol{q}_{(i,j)}$ and vector $\boldsymbol{k}_{(m,n)}$ in the $\boldsymbol{K}$ that are in the same row or column as $\boldsymbol{q}_{(i,j)}$, as follows:

$$\boldsymbol{A}^{i,j}_{(m,n)} = \begin{cases} \frac{\boldsymbol{q}_{(i,j)} \boldsymbol{k}^T_{(m,n)}}{\|\boldsymbol{q}_{(i,j)}\| \|\boldsymbol{k}_{(m,n)}\|} & \text{if } m = i \text{ or } n = j, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

After traversing all vectors in the spatial dimension of $\boldsymbol{Q}$, we obtain $N \times C$ criss-cross attention maps like $\boldsymbol{A}^{i,j}$, denoted by $\boldsymbol{A} = \{\boldsymbol{A}^{1,1}, \boldsymbol{A}^{1,2}, \cdots, \boldsymbol{A}^{N,C}\} \in \mathbb{R}^{N \times C \times (N \times C)}$. Then, we perform the softmax operation over the channel dimention of $\boldsymbol{A}$ to obtain the spatial-temporal attention map, denoted by $\boldsymbol{M} = \{\boldsymbol{M}^{1,1}, \boldsymbol{M}^{1,2}, \cdots, \boldsymbol{M}^{N,C}\}$, as follows:

$$\boldsymbol{M}^{i,j} = \exp\left(\boldsymbol{A}^{i,j}\right) \circledcirc \sum_{m=1}^{N} \sum_{n=1}^{C} \exp\left(\boldsymbol{A}^{m,n}\right), \quad (2)$$

where $\circledcirc$ indicates dividing two matrixes by dividing corresponding elements. $\boldsymbol{M}^{i,j}$ is a sparse matrix with a size of $N \times C$ in which only elements in the i-th row or j-th column are non-zero, representing the correlations between the local spatial-temporal feature $\boldsymbol{F}_{(i,j)}$ and its adjacent spatial-temporal space. We obtain the aggregated features as follows:

$$\tilde{\boldsymbol{F}}_{(i,j)} = \boldsymbol{M}^{i,j} \circledast \boldsymbol{F} + \boldsymbol{F}_{(i,j)}, \quad (3)$$

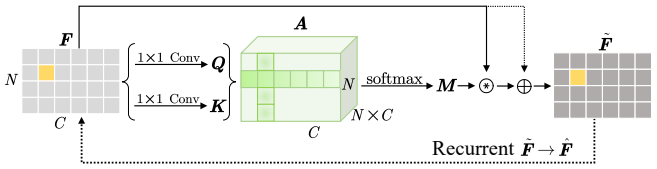

Fig. 2. Details of the spatial-temporal attention.

where $\circledast$ indicates multiplying two matrixes by multiplying corresponding elements. After the first crisscross attention operation, only the connection between pixels in the same row and column has been established. Then we repeat the above process once to establish the connection between any two pixels, denoted by $\tilde{\boldsymbol{F}} \rightarrow \hat{\boldsymbol{F}}$. The STA is easy-to-plugin with a complexity of space and time of $\mathbf{O}((N \times C) \times (N + C))$.

2.3. Training loss

As shown in Figure 1, we denote the boosted features of normal video and abnormal video as $\hat{\boldsymbol{F}}^n = \{\hat{\boldsymbol{f}}_1^n, \cdots, \hat{\boldsymbol{f}}_N^n\}$ and $\hat{\boldsymbol{F}}^p = \{\hat{\boldsymbol{f}}_1^p, \cdots, \hat{\boldsymbol{f}}_N^p\}$, respectively. To enable $\hat{\boldsymbol{F}}^n$ to record the prototypical patterns of normal events while ignoring the diversity, with reference to the cosine embedding loss [16], we introduce a mutual cosine embedding loss, denoted by $\mathcal{L}_{MCE}$, to obtain a more compact feature representation of $\hat{\boldsymbol{F}}^n$ while keeping the features of abnormal clips in $\hat{\boldsymbol{F}}^p$ away as follows:

$$\begin{aligned}\mathcal{L}_{MCE} =& 1 - \underset{1\le i<j\le N}{\text{Avg}} \left(\frac{\hat{\boldsymbol{f}}_i^n \hat{\boldsymbol{f}}_j^n}{\|\hat{\boldsymbol{f}}_i^n\| \|\hat{\boldsymbol{f}}_j^n\|} \right) \\ &+ \underset{1\le i<j\le N}{\text{Avg}} \left(\min \left(\frac{\hat{\boldsymbol{f}}_i^p \hat{\boldsymbol{f}}_j^p}{\|\hat{\boldsymbol{f}}_i^p\| \|\hat{\boldsymbol{f}}_j^p\|}, \xi \right) \right),\end{aligned} \tag{4}$$

where i, j are the order of video clips, and Avg $(\cdot)$ denotes the mean value. The ξ denotes a margin used to ignore normal clips contained in abnormal videos. Following the previous work [2], we apply the MIL ranking loss $\mathcal{L}_{MIL}$ to optimize the bd-RNN-based regression model, as follows:

$$\begin{aligned}\mathcal{L}_{MIL} =& \max \left(0, 1 - \max_{1\le i\le N} r\left(\hat{\boldsymbol{f}}_i^p\right) + \max_{1\le j\le N} r\left(\hat{\boldsymbol{f}}_j^n\right) \right) \\ &+ \lambda_1 \sum_{i=1}^{N-1} \left(r\left(\hat{\boldsymbol{f}}_{i+1}^p\right) - r\left(\hat{\boldsymbol{f}}_i^p\right) \right)^2 + \lambda_2 \sum_{i=1}^{N} r\left(\hat{\boldsymbol{f}}_i^p\right),\end{aligned} \tag{5}$$

where $r(\cdot)$ denote the anomaly score. The first part is ranking loss, used to make the maximum score of instance in the positive bag higher than that in the negative bag. The last two parts are smoothness loss and sparse loss, which are used to encourage the smoothness and sparsity of scores, respectively. Balanced by the λ_{MCE}, the total loss is as follows:

$$\mathcal{L}_{total} = \lambda_{MCE} \mathcal{L}_{MCE} + \mathcal{L}_{MIL}. \tag{6}$$

3. EXPERIMENTS

3.1. Implementation details

Datasets. We conduct experiments on three public benchmarks to evaluate the proposed method. The UCF-crime [2] is a large-scale ws-VAD dataset containing 1900 untrimmed surveillance videos, in which the frame-level labels are only available for testing videos. The ShanghaiTech [17] is a challenging dataset for unsupervised VAD. We follow [12] to reorganize it by moving part of the abnormal videos in the testing set to the training set to adapt to the weakly supervised setting. The UCSD Ped2 [18] is a small-scale and prevailing dataset in the unsupervised VAD task. We randomly select ten normal videos and six abnormal videos as our training set, while the rest are used as the testing set.

Evaluation metrics. Following previous works [2, 8], we use the frame-level area under the curve (AUC) as the main evaluation metric. Besides, we calculate the average score of all negative and positive instances and compare the score gap.

Training details. The model is trained on the PyTorch [19] framework. The pre-processes and segmentation of videos is similar to [2]. The bd-RNN includes two hidden layers with 64 units, and we use the sigmoid activation for the output layer and ReLU for other layers. Adam [20] optimizer is utilized with an initial learning rate of 1×10^{-4}. The trade-off parameter λ_{MCE}, λ_1 and λ_2 are set to $2 \times 10^{-2}, 2 \times 10^{-4}$ and 4×10^{-5}, respectively. We set the number of the channel of query map and key map to 10, and ξ is set to 0.4.

3.2. Quantitative comparison

The quantitative comparison results of frame-level AUC are presented in Table 1. We compare our method with prevailing unsupervised methods [4, 3, 6, 17, 5] and weakly supervised methods [2, 9, 12, 8]. The results demonstrate that our method outperforms state-of-the-art methods on the UCF-crime [2] dataset and achieves the second-best performance on the UCSD Ped2 [18] dataset. From the performance gap between the weakly supervised methods and supervised methods on the ShanghaiTech [17] dataset, ws-VAD is more reliable than unsupervised VAD, which is more suitable to process large-scale, real-world videos. Compared to the weakly supervised methods [2, 9, 8] that directly uses I3D features, our method achieves 5.1%, 4.3%, and 0.7% AUC margins on the UCF-crime [2] dataset, respectively, showing that the feature boosted by the spatial-temporal attention are more discriminative for ws-VAD.

3.3. Ablation studies

To verify the effectiveness of the mutual embedding cosine loss and bd-RNN, we calculate the average anomaly scores with and without $\mathcal{L}_{MCE}$, using bd-RNN or FC. The results presented in Table 2 show that $\mathcal{L}_{MCE}$ can enhance the score gap between positive and negative instances. Compared to

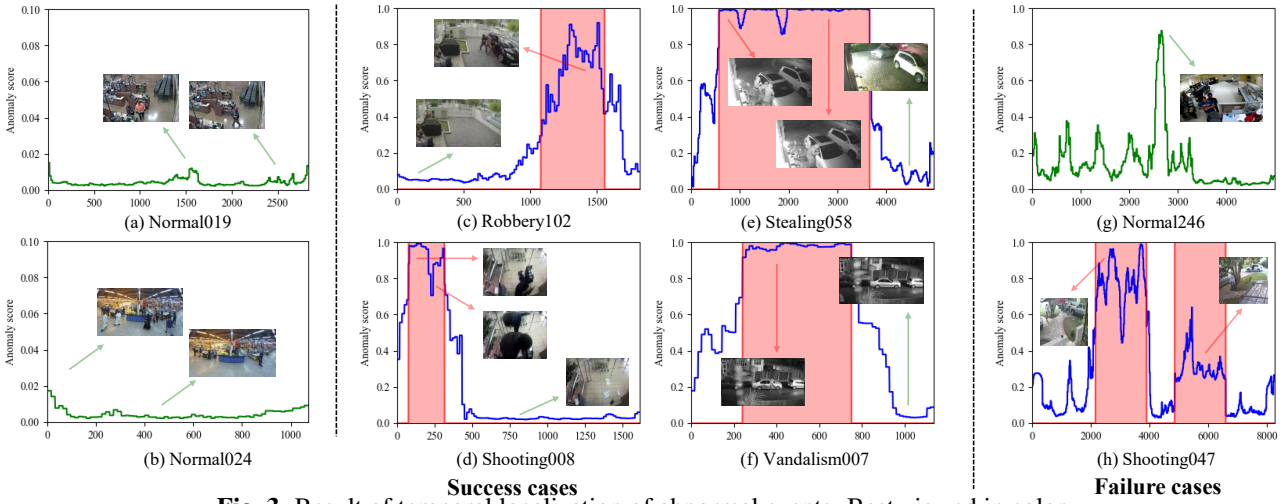

Fig. 3. Result of temporal localization of abnormal events. Best viewed in color.

Table 1. Results of quantitative frame-level AUC comparison. We report the performance of our mehtod on the UCF-crime (UCF), ShanghaiTech (S.T.), and UCSD Ped2 (UCSD) datasets. * indicates the results we reproduce with the corresponding settings. Bold numbers indicate the best performances, and underlined ones indicate the second best.

	Method	Feature	AUC(%)		
			UCF	S.T.†	UCSD‡
Unsupervised	Hassan *et al.* [4]	-	50.6	60.9	-
	Lu *et al.* [3]	-	65.5	-	-
	StackRNN [6]	-	-	68.0	92.2
	Frame-Pred [17]	-	-	72.8	95.4
	Mem-Guided [5]	-	-	70.5	**97.0**
Weakly supervised	Sultani *et al.* [2]	$C3D^{(RGB)}$	75.4	86.3	-
		$I3D^{(RGB)}$	77.9	87.7*	91.8*
	Zhang *et al.*[9]	$I3D^{(RGB)}$	78.7	82.5	-
	Zhong *et al.* [12]	$C3D^{(RGB)}$	81.1	76.4	-
		$TSN^{(RGB)}$	82.1	84.4	-
	MIST [8]	$C3D^{(RGB)}$	81.4	<u>93.1</u>	-
		$I3D^{(RGB)}$	<u>82.3</u>	**94.8**	-
	Ours	$C3D^{(RGB)}$	81.6	88.7	92.3
		$I3D^{(RGB)}$	**83.0**	90.2	<u>96.7</u>

Table 2. Results of ablation studies. We report the average anomaly scores of all negative and positive instances on the UCF-crime dataset with I3D features. ↑ means that a higher value indicates better performance and ↓ vice versa.

Model	Negative ↓	Positive ↑	Gap ↑	AUC (%)
Ours	0.21	0.84	0.63	83.0
Ours w/o $\mathcal{L}_{MCE}$	0.33	0.75	0.42	79.8
Ours w FC	0.24	0.80	0.56	82.2
Ours w FC w/o $\mathcal{L}_{MCE}$	0.35	0.76	0.41	79.4

FC, the bd-RNN can capture the temporal correlations between adjacent clips, resulting in a 0.8% margin in AUC.

3.4. Visual results

Figure 3 shows the results of temporal localization of abnormal events on the UCF-crime [2] dataset with I3D [14] features, where (a)-(f) show successful cases and (g)-(h) show failure cases. As shown in (a) and (b), the scores of all clips in normal videos are close to 0, and the highest scores of clips in videos *normal019* and *normal024* are less than 0.02. In contrast, the scores of abnormal clips in the abnormal videos are close to 1, as shown in the area marked by the red window in (c)-(f). In addition, the average score of abnormal clips is much greater than that of normal clips in the same abnormal videos, demonstrating that the model is able to distinguish normal and anomaly and localize the abnormal behaviors effectively. In the two failure cases, the model gives a high score to normal behaviors in (g) and fails to detect the second shooting in (h). By watching the corresponding raw videos, we find that the surveillance cameras do not fully capture the behaviors due to the improper camera view and shelters.

4. CONCLUSION

In this paper, we propose spatial-temporal attention to capture the correlations between the spatial-temporal features of different video clips and assemble it with a MIL ranking model to obtain the task-specific representation for ws-VAD. Experimental results on three standard benchmarks show that the boosted features are more discriminative than the I3D and C3D features. Our STA augmented MIL ranking model outperforms most of the weakly supervised and unsupervised methods, which can temporally localize abnormal clips efficiently. Future work includes improving the capability to process low-quality videos (low-resolution, occlusion, etc.) and developing online detection methods for real-time detection.

5. REFERENCES

[1] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

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

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

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

[5] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, "Learning memory-guided normality for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14372–14381.

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

[7] Yi Zhu and Shawn Newsam, "Motion-aware feature for improved video anomaly detection," *arXiv preprint arXiv:1907.10211*, 2019.

[8] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng, "Mist: Multiple instance self-training framework for video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14009–14018.

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

[10] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *European Conference on Computer Vision*. Springer, 2020, pp. 322–339.

[11] Charles Bergeron, Jed Zaretzki, Curt Breneman, and Kristin P Bennett, "Multiple instance ranking," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 48–55.

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1237–1246.

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

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

[15] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[16] Yuli Wu, Long Chen, and Dorit Merhof, "Improving pixel embedding learning through intermediate distance regression supervision for instance segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–227.

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

[18] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette, "Real-time anomaly detection and localization in crowded scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 56–62.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.