VIDEO ANOMALY PREDICTION: PROBLEM, DATASET AND METHOD

Yang Wang[1,2], *Jun Xu*[2*], *Jiaogen Zhou*[3], *Jihong Guan*[1*]

[1]Tongji University, [2]Shanghai Bilibili Technology Co., Ltd., Shanghai, China
[3]Huaiyin Normal University, Huaian, Jiangsu, China

ABSTRACT

The task of *Video Anomaly Detection* (VAD) is to find anomalies existing in given videos, which has been extensively studied. This paper addresses the task of Video Anomaly Prediction (VAP), which is to predict whether any anomaly will happen in streaming videos. With VAP, we can intervene or alert before anomalies really occur, thus preventing damage to public life and property. VAP is a more significant yet challenging task, which currently has only one work in the literature. To challenge this task, we first propose the concepts of predictable and unpredictable anomalies, and define the VAP task in video surveillance scenario. Based on this definition, we then construct the first VAP dataset, which consists of 618 frame-level annotated videos, including 300 normal videos and 318 anomaly videos. Next, we propose a *Dual-channel Video Anomaly Prediction* (DVAP) method based on feature enhancement and a new annotation scheme. Finally, We evaluate the proposed method and compare it with related works from classification and regression perspectives. Experimental results show that DVAP can predict anomalies one second earlier than they really occur and achieves the best accuracy.

***Index Terms*—** Video anomaly Prediction; Dataset; Dual-channel; Classification; Regression.

1. INTRODUCTION

Video surveillance is a core tool for detecting anomalous events in the field of public safety management. Using computer vision (CV) technology to automatically analyze video content can significantly improve the effectiveness and efficiency of anomaly detection. Therefore, video anomaly detection (VAD) has become a hot research task in the area of CV. The major goal of VAD is to discover anomalies and their locations (usually timestamps) in the given videos [1]. Currently, there are mainly two types of methods for VAD: unsupervised and weakly supervised. Unsupervised methods assume that there are only a few number of anomalies in each given video, and exploit the observation that the reconstruction/prediction errors of abnormal frames are usually large for anomaly detection [2, 3, 4, 5, 6, 7]. Weakly supervised methods using only video-level annotations are mainly based on a multi-instance learning framework and can produce more accurate frame-level anomaly scores than unsupervised methods [8, 9, 10, 11, 12]. Although existing VAD methods can detect anomalies effectively, they cannot provide early predictions of anomalies for streaming videos. The reason is twofold: 1) The video anomaly prediction (VAP) task is not properly defined; 2) Existing datasets for VAD contain a large number of unpredictable anomalies, which are not suitable for the research of VAP. Recently, Lent et al. proposed an unsupervised Ex-ante Potential Anomaly Prediction Network for anomaly warning [5, 6]. However, it can only predict up to five frames, is not effective for the VAP task.

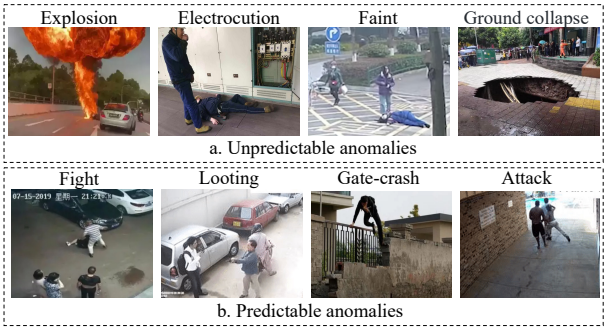

Fig. 1. Examples of predictable and unpredictable anomalies.

In this paper, we consider the VAP task, which tries to determine whether any anomaly will happen based on the known video content in surveillance scenario. To this end, we give a formal definition of the VAP task, and construct the first VAP dataset. Furthermore, we develop a Dual-channel Video Anomaly Prediction (DVAP) method based on feature enhancement. The feature enhancement module enhances the representations of key features in videos. This dual-channel structure divides anomaly prediction into two parallel subtasks: distinguishing anomaly video states and distinguishing normals from anomalies in video content. The effective anomaly prediction is achieved by merging the results of the two subtasks. Additionally, we design a more suitable annotation scheme for video anomaly prediction. We also implement the method from both classification and regression perspectives.

In summary, the main contributions of this paper are: 1) We give a formal definition of the video anomaly prediction task and construct the first VDP dataset. 2) We propose a dual-channel VAP (DVAP) method based on feature enhancement and a new annotation scheme. 3) We conduct extensive experiments on the constructed dataset, and compare DVAP with existing models from both classification and regression perspectives. Experimental results show that our method outperforms the existing models and succeeds in anomaly prediction one second in advance.

2. PROBLEM STATEMENT

We consider video anomaly prediction (VAP) in video surveillance scenario as a task that is to predict possible anomalies before the anomalous events really occur. Therefore, with VAP we can alert and intervene to prevent damage to human life and property, making

*Corresponding authors: zijun@bilibili.com, jhguan@tongji.edu.cn.
This work was supported in part by National Key R&D Program of China (No. 2021YFC3300300) and National Natural Science Foundation of China (No. U1936205).

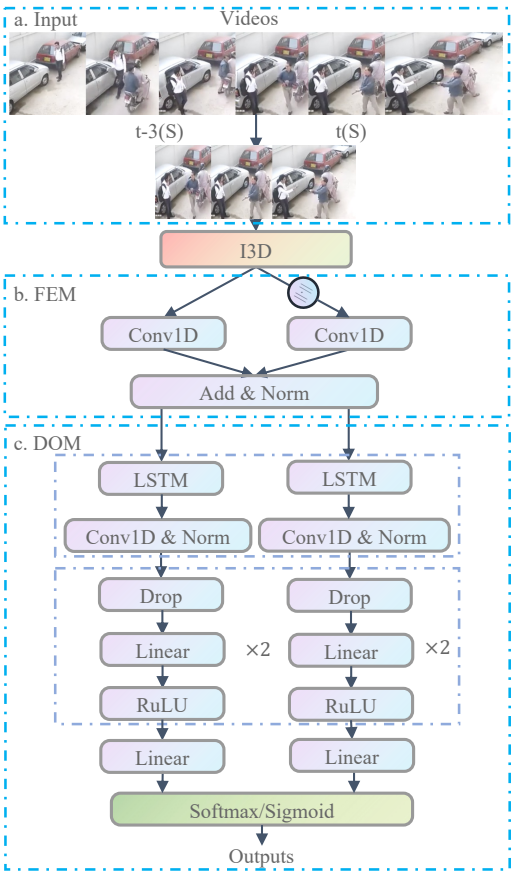

Fig. 2. The framework of our method DVAP.

it important in public safety and also a challenging research problem. To challenge this problem, it is essential to define what anomalies can be predicted. Hence, we divide anomalies into two categories: 1) predictable anomalies, which exhibit visual cues before their occurring and can be captured by analyzing the videos; and 2) unpredictable anomalies, which either lack visual cues before their occurrence or cannot be computationally captured from the videos. Among them, predictable anomalies form the core for dataset construction and method development, while unpredictable anomalies will not be discussed further. Fig. 1 shows some examples of the two types of anomalies. Additionally, in Sec. 5.2 of this paper, we will validate the rationale of our anomaly categorization through designed experiments.

3. DATASET

As there is no dataset specifically for the VAP task, we construct one from scratch. Through keyword search, we obtain nearly 20,000 abnormal videos from platforms including Bilibili, YouTube, and the UCF-Crime dataset. After screening, we acquire 318 abnormal surveillance videos and 300 normal surveillance videos that meet our criteria. Among them, there are 218 predictable abnormal videos, mainly including the categories of unauthorized intrusion, fighting, robbery, arson, assault, and vandalism; and 100 unpredictable abnormal videos, mainly including the categories of explosions, electric shocks, sudden illnesses, natural disasters, and sudden accidents.

There are two annotation schemes used for this dataset. One is video-level annotation, which classifies the videos into two categories: normal and abnormal. The other is frame-level annotation, specifically designed for regression-based video anomaly prediction. The specific details will be introduced in the method section.

4. METHOD

As shown in Fig. 2, our DVAP method consists of an input module (Input), a feature enhancement module (FEM), and a dual-channel output module (DOM). In the input module, considering the short interval between the beginning of the video and the occurrences of the anomalies in the dataset, we selectively extract a clip of t (set to 3 in this paper) seconds from each video as an input sample.

In normal videos, we randomly clip a t-second video. For anomalous videos, we adopt two clipping methods. The first method selects the video from the t-th second before the anomaly to the moment of anomaly occurrence, representing the feature of the entire anomaly occurrence. The second method selects the video from the $(t+x)$-th second before the anomaly to the x-th second before the anomaly. The video selected by the second method helps the model understand the video features of the x seconds' video content before the anomaly, also referred to as the "prediction video". The video features are input into the pre-trained I3D [13], which provides inputs for FEM, outputting enhanced video features. Finally, the task is decomposed via DOM into two channels: one channel outputs the state of the predicted anomaly (how much time remains before the anomaly occurs), and the other determines if an anomaly will occur. By integrating the results of both channels, VAP is done.

4.1. Feature Enhancement

In FEM, we obtain the input feature matrix F from I3D, where $F = \{f^{b,n,t,c} | b \in [1,B], n \in [1,N], t \in [1,T], c \in [1,C]\} \in \mathbb{R}^{B \times N \times D \times C}$. Here, B denotes batch size, N represents the number of iterations for feature enhancement [8], while T and C indicate that the input video feature is equally divided into T clips, with each clip having a feature dimension of C. Along the C dimension, we calculate the Euclidean norm to obtain the feature variation matrix $W^{B,N,T}$, which represents the intensity of the overall information of the video feature matrix. The detailed implementation is as follows:

$$W^{B,N,T} = \left\| f^{b,n,t,c} \right\|_2 \tag{1}$$

Next, we utilize $conv1d$ to perform feature interaction of feature matrix F in the T dimension. This helps the model in capturing the feature correlation between adjacent clips in the video. Similarly, we adjust the feature dimension of W through $conv1d$, which is used to enhance video feature F. The enhanced video feature (F_{FEM}) is obtained as follows:

$$F_{FEM} = Conv1D(F) + \alpha(Conv1D(W^{B,N,T})) \tag{2}$$

wherein, α is a hyperparameter used to control the degree of enhancement.

4.2. Dual-channel Prediction

We propose an innovative task-decomposition dual-channel output module to accomplish the highly complex task of VAP. The structure of this module is shown in Part 'c' of Fig. 2. The composition of the two channels is consistent. The features are initially passed through an LSTM [14] layer to enhance the model's extraction of

overall temporal features. Then, the features go through stacked linear layers containing $Drop$ and $ReLU$ to help the model capture complex patterns in the features more effectively and prevent overfitting. Finally, to more comprehensively explore the implementation of VAP, we also investigate the prediction effect of the model from both classification and regression perspectives. Therefore, we use $Softmax$ and $Sigmoid$ in the output layer to produce different forms of output respectively.

4.2.1. Prediction by Classification

When treating VAP as a classification problem, we decompose the task into two subtasks: 1) distinguishing the state of the anomaly videos, and 2) distinguishing between normal and anomaly videos. Sub-task 1 inputs videos where an anomaly has just occurred and videos from the x-th second before the anomaly, while sub-task 2 takes normal videos and videos where an anomaly has just occurred as input samples. The output layer of this module uses the $Softmax$ function and optimizes the model using cross-entropy loss. The training strategy is to first train sub-task 1, then train sub-task 2 with sub-task 1's parameters fixed, to obtain the final VAP model. The loss function is as follows:

$$Loss_{cls} = -\sum_{i=1}^{SN}\sum_{j=1}^{SC} y_{ij}\log(SoftMax(f_i)_j) \tag{3}$$

where $Loss_{cls}$ is the classification loss, SN is the number of samples, SC represents the number of classes, y_{ij} indicates whether the i-th sample belongs to the j-th class, and $Softmax(f_i)^j$ represents the probability that the $Softmax$ function predicts the final feature f_i of the i-th sample as the j-th class.

4.2.2. Prediction by Regression

By treating VAP as a regression problem, we can use the existing annotation methods [8] to assign the labels '0' and '1' to normal and abnormal frames, respectively. However, these methods cannot differentiate various abnormal states. Therefore, we design a new annotation scheme, aiming to more delicately annotate normal and different abnormal states for regression.

In VAD, it is assumed that the anomaly scores of adjacent video clips are continuous [8]. Psychological studies indicate that individuals tend to take more proactive or significant measures as they approach a goal [15]. Based on this insight, we assume that the scores of predictable anomalous videos follow a normal distribution, closely reaching 1 when an anomaly occurs. With this assumption, we present the scoring curves for normals, anomalies, and prediction videos, as depicted in Fig. 3. Specifically, (a) represents the standard scoring curve of T (set to 32) clips of a t-second (set to 3) normal video that are uniformly clipped. Similarly, (b) shows the standard scoring curve of the abnormal video from the t-th second before the anomaly until the moment it occurs, with the score ranging from 0 to 1 and fitting a normal distribution. (c) illustrates the standard scoring curve of the prediction video from the (t+1)-th second before the anomaly to 1 second prior, with the score ranging from 0 to 0.607, encompassing the first 2/3 of the (b) case and the 1/3 standard score for the (a) case.

Based on the aforementioned annotation scheme, we re-annotate the dataset and retrain the model. The model framework and training strategy of the regression-based dual-channel output module remain unchanged. However, in the output layer, we adopt the $Sigmoid$ function and have each sample continuously output 32 scores. For

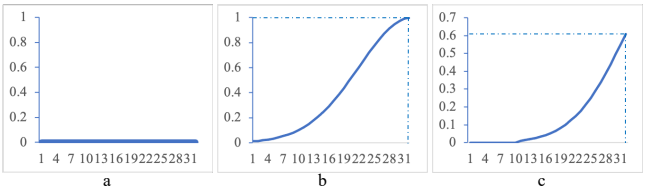

Fig. 3. Abnormal scoring curves for three types of video samples.

the two decomposed sub-tasks, we determine their results based on the average predicted scores of the last three clips of the samples. Finally, the loss of the model combine the MSE loss and the continuity loss. The overall training loss function is as follows:

$$Loss_{Reg} = \frac{1}{SN}\sum_{i=1}^{SN}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{SN}\sum_{j=1}^{T-1}|\hat{y}_{i,j+1} - \hat{y}_{i,j}| \tag{4}$$

Above, the left half represents the MSE loss, which constrains the predicted value to be as consistent as possible with the true label. The right half represents the continuity loss, which constrains the difference between adjacent scores to approach zero, making the score more stable. Here, y_i is the value of the standard scoring curve of the sample, $\hat{y}_i$ is the value of the predicted scoring curve of the sample, and $\hat{y}_{i,j}$ is the score of the j-th clip of sample i.

5. PERFORMANCE EVALUATION

5.1. Experimental Details

We utilize the I3D network pre-trained on the Kinetics-400 [16] dataset for feature extraction, which can extract video samples into 32 clips, each with a continuous feature of length 2048. The training process is divided into two stages. First, we focus on distinguishing the abnormal states videos. Then fixing the parameters of the first stage, we proceed to distinguish between normal and abnormal videos. We use the Adam optimizer, with a learning rate set to 0.001, a weight decay coefficient of 0.0005, a batch size of 64, the hyperparameter α is set to 0.1, and a total training duration of 300 epochs. Moreover, the model is implemented in the PyTorch framework and is trained using an RTX 3090. The evaluation metric is accuracy (acc), which is a commonly used evaluation criterion, often employed to measure the overall accuracy of a model. It is defined as the ratio of the samples correctly predicted by the model to the total number of samples.

Method	A (acc%)	B (acc%)
TSN_r50 [17] (ECCV 2016)	86.1	**52.2**
TSM [18] (ICCV 2019)	83.2	51.8
TimeSformer [19] (ICML 2021)	87.1	49.2
VideoMAE2 [20] (CVPR 2023)	**86.9**	50.4

Table 1. Abnormal vs. normal video classification experiment.

5.2. Results

The anomalies are divided into "predictable anomalies" and "unpredictable anomalies". To verify this definition, we extract 100 clips (from the (t+1)-th second, to the 1-th second before the anomaly) from each type of anomalous videos and train them against normal videos using both the most classic and the latest video classification models. As shown in Table 1, experimental results from classification models indicate that the distinction between predictable

Method	Normal and Anomaly videos (acc%)	Normal and 4 Anomaly states (acc%)	Normal and 2 Anomaly states (acc%)	4 Anomaly States (acc%)	2 Anomaly states (acc%)	Dual-channel (acc%)
TSN_r50 [17]	86.7	20.2	40.7	13.7	55.6	-
TSM [18]	83.7	13.3	38.6	11.1	55.3	-
TimeSformer [19]	86.7	16.7	42.4	12.7	50.2	-
VideoMAE2 [20]	**87.4**	16.7	41.8	5.8	55.4	-
DVAP(Cls)	86.8	**22.1**	**43.4**	**31.2**	**70.0**	**85.3/70.0**

Table 2. Classification Results. "4 Anomaly states" refers to videos of 4 states: the moment of anomaly, and 0.1, 0.5, and 1 second before the anomaly occurs. "2 Anomaly states" refers to videos of two states: the moment of anomaly and 1 second before the anomaly occurs. "Dual-channel" indicates the results of our dual-channel task decomposition approach.

anomaly videos and normal videos has an accuracy rate of over 85% *(A)*. In contrast, unpredictable anomaly videos and normal videos cannot be distinguished *(B)*. The results suggest that unpredictable anomaly videos cannot be distinguished from normal videos before the anomaly occurs, making unpredictable anomaly videos unsuitable for VAP.

5.2.1. Classification Results

To examine anomaly prediction from a classification perspective, we compare DVAP with existing video classification methods under various experimental conditions. As indicated in Table 2, we conclude: 1) Normal and anomalous videos can be distinguished clearly, but differentiating specific anomaly states is challenging. The major reason lies in that the feature differences between various anomaly states are too subtle. 2) Most models cannot effectively differentiate anomaly states. Only DVAP achieves 70.0% accuracy in classifying imminent anomalies and states a second before the anomaly occurs. Therefore, future experiments will focus on predictions made a second in advance. 3) Combining conclusions 1 and 2, we can see that it is challenging to achieve anomaly prediction with a single task. Therefore, we propose a dual-channel task decomposition strategy, which is the first effective prediction strategy, with accuracy of 70.0% and 85.3% for sub-tasks 1 and 2, respectively.

5.2.2. Regression Results

From a regression perspective, we further explore VAP. First, we propose a new annotation method and re-annotate the dataset. Then, we conduct experiments on this dataset and compare it with the classification-based VAP. We determine the prediction results based on the average predicted scores of the last three clips of the video samples. For the distinguishing between normal and abnormal subtask, we set a threshold of 0.5. Samples with an average prediction score exceeding the threshold are judged as anomaly videos. For the sub-task of distinguishing abnormal states, the threshold is set to 0.75. Samples with an average prediction score exceeding this threshold are judged as videos where an anomaly has already occurred. Finally, according to the results in Table 3, the regression-based DVAP model has the best accuracy in all three experiments. This indicates that the new annotation method is a more detailed and effective approach, making it more suitable for the VAP task. The VAD model [9, 12] using the 0/1 annotation method in the VAP problem has an extreme data imbalance problem and therefore cannot be compared. Training is not effective.

Method	A (acc%)	B (acc%)	Dual-channel (acc%)
DVAP (Cls)	86.8	70.0	85.3/70.0
DVAP (Reg)	**90.2**	**71.1**	**88.8/71.1**

Table 3. Regression results. 'A' and 'B' respectively indicate the results of 'Normal and Anomaly videos' and '2 Anomaly states'.

Index	FEM	DOM	ALM	Dual-channel (acc%)
A	-	-	-	83.4/56.6
B	√	-	-	84.2/66.7
C	√	√	-	85.3/70.0
D	√	√	√	**88.8/71.1**

Table 4. Ablation results

5.3. Ablation Study

We conduct ablation studies to assess the effects of the Feature Enhancement Module (FEM), the temporal feature extraction part of the Dual-channel Output Module (DOM), and the anomaly annotation method based on the normal distribution assumption (ALM). As shown in Table 4, FEM enhances the differences between features, thus significantly improves the distinction accuracy of anomaly states under low feature difference situations *('A' vs. 'B')*. DOM further optimizes the overall task accuracy *('B' vs. 'C')*, highlighting the critical importance of video temporal information. Finally, based on the results of ALM, we can see that ALM is a more ideal annotation strategy for the current problem. It allows the model to have a deeper understanding of the training data, hence achieving the best accuracy *('C' vs. 'D')*.

6. CONCLUSION

This paper formally defines the task of Video Anomaly Prediction (VAP) for the first time, and constructs a dedicated dataset for the VAP task. Furthermore, we propose a dual-channel video anomaly prediction method DAVP based on feature enhancement and a more suitable video annotation scheme. We conduct extensive experiments on the dataset to evaluate DVAP. Based on our experiments, we show that anomalies in surveillance videos can be categorized into "predictable" and "unpredictable" types. Experimental results indicate that currently, only our model is capable of effectively predicting anomalies one second in advance, which is significantly important in the field of public safety. However, there is still much room for improvement in this research. In the future, we will explore methods to better understand the behavioral information among individuals in open video scenes, aiming to enhance prediction performance and optimize video prediction strategies for different prediction times and types of anomalies.

7. REFERENCES

[1] Bharathkumar Ramachandra, Michael J. Jones, and Ranga Raju Vatsavai, "A survey of single-scene video anomaly detection," *TPAMI*, vol. 44, no. 5, pp. 2293–2312, 2022.

[2] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 2019, pp. 1705–1714, IEEE.

[3] Hyunjong Park, Jongyoun Noh, and Bumsub Ham, "Learning memory-guided normality for anomaly detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. 2020, pp. 14360–14369, Computer Vision Foundation / IEEE.

[4] Guoqiu Li, Shengjie Chen, Yujiu Yang, and Zhenhua Guo, "A two-branch network for video anomaly detection with spatio-temporal feature learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao, "Future frame prediction for anomaly detection - A new baseline," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 2018, pp. 6536–6545, Computer Vision Foundation / IEEE Computer Society.

[6] Jiaxu Leng, Mingpi Tan, Xinbo Gao, Wen Lu, and Zongyi Xu, "Anomaly warning: Learning and memorizing future semantic patterns for unsupervised ex-ante potential anomaly prediction," in *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. 2022, pp. 6746–6754, ACM.

[7] Shengyang Sun and Xiaojin Gong, "Hierarchical semantic contrast for scene-aware video anomaly detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. 2023, pp. 22846–22856, IEEE.

[8] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 2018, pp. 6479–6488, Computer Vision Foundation / IEEE Computer Society.

[9] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. 2021, pp. 4955–4966, IEEE.

[10] Yang Xiao, Liejun Wang, Tongguan Wang, and Huicheng Lai, "Scoreformer: Score fusion-based transformers for weakly-supervised violence detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang, "Unbiased multiple instance learning for weakly supervised video anomaly detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. 2023, pp. 8022–8031, IEEE.

[12] Yingxian Chen, Zhengzhe Liu, Baoheng Zhang, Wilton W. T. Fok, Xiaojuan Qi, and Yik-Chung Wu, "MGFN: magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in *Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville, Eds. 2023, pp. 387–395, AAAI Press.

[13] João Carreira and Andrew Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 4724–4733, IEEE Computer Society.

[14] Alex Graves and Alex Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[15] Andrea Bonezzi, C Miguel Brendl, and Matteo De Angelis, "Stuck in the middle: The psychophysics of goal pursuit," *Psychological science*, vol. 22, no. 5, pp. 607–612, 2011.

[16] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017.

[17] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*. 2016, vol. 9912 of *Lecture Notes in Computer Science*, pp. 20–36, Springer.

[18] Ji Lin, Chuang Gan, and Song Han, "TSM: temporal shift module for efficient video understanding," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 2019, pp. 7082–7092, IEEE.

[19] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 813–824, PMLR.

[20] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao, "Videomae V2: scaling video masked autoencoders with dual masking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. 2023, pp. 14549–14560, IEEE.