Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: 10.61784/jcsee3005

VIDEO ANOMALOUS BEHAVIOUR DETECTION BASED ON COMPRESSED-INFLATED ATTENTION MODULE

DengBin Xu[1], PeiChen Wu[1], LiNing Yuan[2*]
[1] *School of Information Network Security, People's Public Security University of China, Beijing 100038, China.*
[2] *School of Public Security Big Data Modern Industry, Guangxi Police College, Nanning 530028, Guangxi, China.*
Corresponding Author: LiNing Yuan, Email: yuanlining@gxjcxy.edu.cn

Abstract: The performance of current attention based feature fusion methods depends on the correlation between features. After feature fusion, due to the inter domain differences of different features, the spatiotemporal perception ability is insufficient, and effectively fusing two cross domain features still faces challenges. A video anomaly detection method based on compression inflation attention feature fusion is proposed to address the issues of insufficient cross domain expression ability of RGB features and optical flow features, as well as weak spatiotemporal perception ability of fused features. The use of Squeeze and Inflation Networks (SENet) to construct a fusion mechanism for RGB and optical flow features can enhance the expression ability of fused features while reducing the number of network parameters and improving the performance of anomaly detection algorithms. In the global spatiotemporal awareness stage, the ConvLSTM (Long Short Term Memory Convolutional Network) is used to achieve global spatiotemporal awareness, while balancing computational complexity and detection performance. We achieved a recognition performance of 93.72% on the UCSDPed2 dataset and also performed well on the CUHKAvenue and LAD2000 datasets, verifying the effectiveness of the method.
Keywords: Computer vision; Abnormal behaviour detection; Feature fusion; Attentional mechanism; Multi-branch convolution

1 INTRODUCTION

With the advent of the information age, human society has made significant progress but also faces new challenges. Currently, social structures are increasingly polarized. Although social conflicts have eased compared to several decades ago, the diversity and complexity of social risks have grown with time. It is clear that traditional methods are insufficient to effectively address these new problems. To ensure the safety of people's health and property, electronic surveillance devices are widely used in public places. For instance, most public areas, such as banks, campuses, parks, and streets, are now equipped with cameras. However, traditional surveillance cameras primarily record video information and lack the capability to automatically detect and alert abnormal behaviors. As a result, these conventional systems can no longer meet the demands of contemporary society [1].

The optimal solution to this issue is to design a high-performance video anomaly detection system. This system can monitor targets in real-time, automatically detect abnormal behaviors in the video, and issue alerts, thereby eliminating the need for manual inspections and managing multiple cameras. This approach not only enhances the intelligence of surveillance systems but also significantly saves human and material resources.

Video Abnormal Behaviour Detection is a technique that analyses whether the target and its motion state in the video data stream are normal or not. With the rapid development of society and economy, public security issues are becoming increasingly complex. Traditional manual surveillance methods can no longer meet the needs of large-scale and long-time surveillance, and the popularity of video surveillance equipment has brought about massive amounts of video data. The application of abnormal behaviour detection technology in these video surveillance systems can quickly and effectively identify abnormal events such as traffic accidents, fights, explosions, etc., and provide early warning signals, so as to protect the lives and properties of the people.

The abnormal behaviour detection algorithms proposed in recent years include unsupervised learning methods, weakly supervised learning methods and fully supervised learning methods. Unsupervised methods learn normal behavioural patterns in the training phase and judge video frames that do not conform to normal behavioural patterns as abnormal [2-4]; weakly supervised learning methods use video-level coarse-grained labelling information in the training phase [5-7] and fully supervised learning methods use frame-level fine-grained labelling information in the training phase [8]. In video feature extraction, spatial and temporal features are usually extracted from the video, and since they belong to different domains, how to achieve effective cross-domain feature fusion is the key to improve the detection results [9]. Commonly used feature fusion includes feature fusion network, weighted average fusion or feature combination approach [8-12].

In recent years, good results have been achieved in feature fusion by learning weight distribution through attention mechanisms, such as Multi-Head Self-Attention (MHSA) [11], Multi-scale Channel Attention Module（MS-CAM) [13], Attention Feature Fusion(AFF) [13]. Methods based on multi-branch aggregation [14] While extracting the time-series motion features of normal events, Transformer is used to achieve multi-layer feature fusion and improve the feature

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

optimisation capability of the encoder. Self-encoder based methods adopt Residual Time Shift Module and Residual Channel Attention Module to enhance the network's ability to model temporal information and channel information respectively [15]. Squeeze-and-Excitation Networks (SENet) [10] is a mechanism to improve Convolutional Neural Networks (CNNs), which enhances the representation of the network by re-weighting the features of each channel, and achieves adaptive adjustment of the weights of each feature channel through three steps: compression, excitation, and relabelling. Applying SENet to the problem of anomalous behaviour detection allows better integration of features from both RGB and optical flow inputs, thus improving the performance of the model. Meanwhile the fused features are usually kept spatio-temporally consistent using a single convolution method [8]. Although the feature expression ability of a single convolution is weak, it can reduce the computational complexity while effectively guaranteeing the performance of the model.

In summary, this paper proposes a video anomalous behaviour detection method based on feature fusion of compression-inflation attention module. In the feature fusion stage, a new mechanism of cross-domain fusion of features is constructed through the SENet module, and the extracted features are fused across domains to effectively improve the accuracy of abnormal behaviour detection. In the spatio-temporal perception stage the spatio-temporal consistency of the fused features is maintained through the ConvLSTM module, and finally the anomaly scores are obtained through the fully connected layer to enhance the spatio-temporal perception of the fused features and to achieve multi-branch and multi-scale feature extraction in the training stage while keeping the high efficiency in the inference stage unchanged, so as to further improve the detection effect.

2 RELATED WORK

2.1 Multi-Head Self-Attention Mechanism

The multi-head self-attention mechanism was first utilized in the field of Natural Language Processing (NLP) and has since been widely applied to various tasks, achieving notable results. This mechanism computes weighted attention scores for each feature through multiple parallel attention computations, allowing it to attend to features at different positions. For feature extraction from video frame sequences, the process begins by initializing the query vector (Q), key vector (K), and value vector (V), and then using multiple (h) attention heads for training. This set of matrices can be used to map video frame features to different subspaces [16].

However, the original multi-head self-attention mechanism did not account for positional information, necessitating position embedding operations on the preprocessed video frame data. This ensures that the positions of the video sequence corresponding to the input video frames are encoded with specific positional information [11].

2.2 Multi-Scale Channel Attention Module

The multi-scale channel attention module (MS-CAM) enhances the model's representation and classification performance by weighting features at different scales. Specifically, MS-CAM is inspired by the ParseNet [17] concept and combines local and global features of convolutional neural networks (CNNs). It utilizes attention modules in the spatial domain to integrate multi-scale information, thereby improving the model's ability to capture and leverage features at various levels of granularity [13].

2.3 Attention Feature Fusion

Attentional Feature Fusion (AFF) [13], based on the Multi-Scale Channel Attention Module (MS-CAM), features a relatively simple structure that employs two branches of different scales to extract channel attention weights. One branch uses Global Average Pooling to capture the attention of global features, while the other branch directly uses point-wise convolution to extract the channel attention of local features.

When fusing two given features X and Y (with Y representing features with a larger receptive field), AFF first performs an initial feature fusion on X and Y. The resulting initial features are then processed through the MS-CAM module. After applying the Sigmoid activation function, the output values range between 0 and 1. To achieve a weighted average of X and Y, the authors use 1 minus the fusion weights, enabling soft selection. Through training, the network automatically determines the weights for each feature.

3 MODELS AND ALGORITHMS

3.1 Algorithmic Framework

The network structure of the anomalous behaviour detection algorithm proposed in this paper is shown in Figure 1 and is divided into four substructures: feature extraction, feature fusion, global spatio-temporal awareness and anomalous score prediction. The input video is decomposed into N video segments, where each video segment contains m frames, and features on RGB and optical flow inputs are extracted on each video segment. The features are passed through an attention-based feature fusion module, SENet, and fed into a global spatio-temporal perception module to obtain a global spatio-temporal feature map, which predicts the video frame anomaly scores through a fully connected layer.

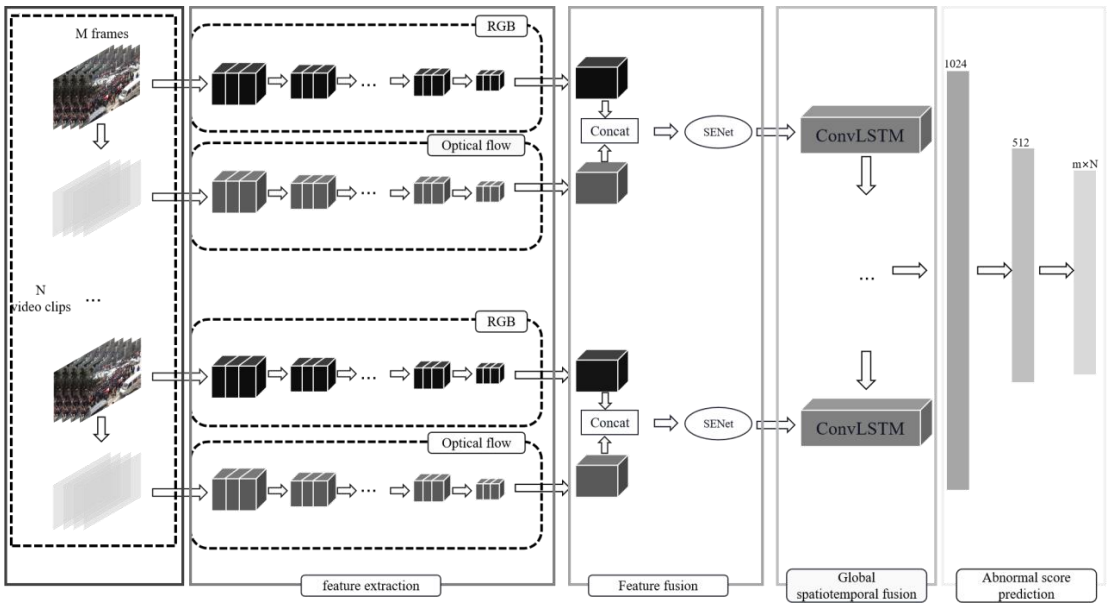

Figure 1 Framework of Anomaly Detection Algorithm

3.2 Attention-based Feature Fusion Module

In the feature extraction phase, features on the RGB input are mainly used to describe the colour information of individual video frames in a video sequence, and are suitable for describing spatial and appearance information in static scenes. The features on the optical flow input, on the other hand, are used to infer the motion by analysing the pixel changes between consecutive frames and represent the motion information in a dynamic scene. The features obtained from RGB and optical flow describe video data from different perspectives and belong to different domains. How to effectively fuse the two so as to improve the effectiveness of video feature representation is a key issue to improve the performance of abnormal behaviour detection algorithms. First, the video is split into video clips with the same number of frames. Then each segment is input into the pre-trained I3D backbone network, and the spatial features of the video sequence are extracted by using the convolution layer and pooling layer to perform multiple convolution and pooling operations on the input data, and a set of feature maps are outputted for each video segment, thus obtaining RGB local features and optical flow local features.

In the feature fusion stage, in order to achieve the cross-domain fusion of two features, this paper adopts an attention-based feature fusion module SENet, which ensures the fusion effect while reducing the computational complexity. The network structure is shown in Figure 2.

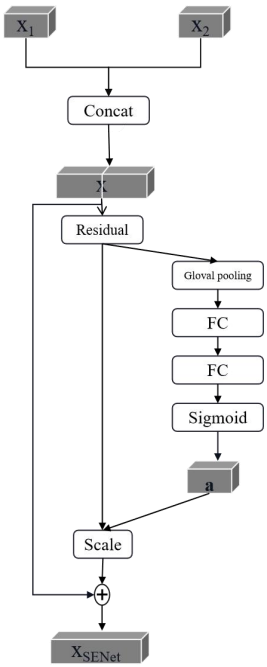

Figure 2 Network Structure Diagram of Squeeze-and-Excitation Networks

In the specific implementation, the input video frame size is scaled to 224×224, and after I3D feature extraction, each video clip gets RGB and optical flow local features of dimension 1024, notated as x_1 and x_2, respectively.
The extracted features x1 and x2 are stacked together as x with a dimension of 5 × 60 × 2048, denoted as $x = (x_1, x_2)$.

$$\boldsymbol{x} = \text{concat}(\boldsymbol{x}_1, \boldsymbol{x}_2) \quad (1)$$

The Squeeze operation is implemented by global average pooling after passing x through the residual layer so that the second and third dimensions of x are normalised to 1. Subsequently, x is input to the two fully connected layers to model the correlation between the channels and the weight tensor describing the similarity is obtained by outputting the same number of weights as the input features, which is normalised as a by Sigmoid.

$$\boldsymbol{a} = Sigmoid(\text{Linear}(\boldsymbol{x})) \quad (2)$$

Linear in the above equation denotes the fully connected layer and the resulting vector $\boldsymbol{a}$ is of size 5 × 1 × 1. Finally the normalised weights are weighted to the features of each channel by the Scale operation. The result is obtained by doing the vector product of x and a and then summed with x to obtain the fusion feature X_{SENet}.

$$\boldsymbol{X}_{SENet} = \boldsymbol{x} + (\boldsymbol{a} \otimes \boldsymbol{x}) \quad (3)$$

The fusion feature X_{SENet} is obtained which can be regarded as a weighted sum of RGB and optical flow features, and the weights are determined by the probability distribution of cosine similarity. This fusion method can make reasonable use of the similarity between RGB and optical flow features to effectively achieve cross-domain fusion, thus improving the robustness of the model and reducing the computational complexity by using feature complementarity.

3.3 Global Spatial and Temporal Awareness

Since video sequences are high-dimensional data, it is crucial for learning an effective anomaly detection model that retains the important information in the time series and filters out the redundant information, so after feature fusion, the fused features of K consecutive segments are fed into a two-layer long and short-term memory convolutional network (ConvLSTM) [18] to achieve global spatio-temporal sensing.
The formulas for ConvLSTM are shown as follows:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{x_i} * \boldsymbol{X}_t + \boldsymbol{W}_{h_i} * \boldsymbol{H}_{t-1} + \boldsymbol{W}_{c_i} \circ \boldsymbol{C}_{t-1} + \boldsymbol{b}_i) \quad (4)$$

$$o_t = \sigma(\boldsymbol{W}_{x_o} * \boldsymbol{X}_t + \boldsymbol{W}_{h_o} * \boldsymbol{H}_{t-1} + \boldsymbol{W}_{c_o} \circ \boldsymbol{C}_t + \boldsymbol{b}_o) \quad (5)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{x_f} * \boldsymbol{X}_t + \boldsymbol{W}_{h_f} * \boldsymbol{H}_{t-1} + \boldsymbol{X}_{c_f} \circ \boldsymbol{C}_{t-1} + \boldsymbol{b}_f) \quad (6)$$

where, $\boldsymbol{X}_t$ denotes the input data x at the current moment t, $\boldsymbol{H}_{t-1}$ denotes the hidden state at the previous moment, $\boldsymbol{W}$ and $\boldsymbol{b}$ denote the weight and bias, respectively, and σ denotes the sigmoid function. $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{o}_t$ denote the output values of the input gate, the forget gate, and the output gate:

$$\boldsymbol{C}_t = \boldsymbol{f}_t \circ \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \circ \tanh(\boldsymbol{W}_{x_c} * \boldsymbol{X}_t + \boldsymbol{X}_{h_c} * \boldsymbol{H}_{t-1} + \boldsymbol{b}_c) \quad (7)$$

$$\boldsymbol{H}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{C}_t) \quad (8)$$

where, ◦ denotes the element-by-element multiplication, tanh denotes the hyperbolic tangent function, $\boldsymbol{C}_t$ and $\boldsymbol{H}_t$ denote the cell state and the hidden state at the current moment, $\boldsymbol{W}_c$ and $\boldsymbol{b}_c$ denote the weight and bias of the cell state, respectively, and $\boldsymbol{H}_t$ represents the output of the ConvLSTM, the global spatiotemporal-aware output result.

3.4 Loss Function

In order to minimise the difference between the predicted anomaly scores and the anomaly labels, the anomalous behaviour detection task is treated as a regression problem, and the L2 distance between the two is calculated as the loss function, i.e., the smoothing loss [19]. The loss function calculation formula is shown as follows:

$$L2 = \sum_i (smooth(s_i - \hat{s}_i)) \quad (9)$$

$$smooth(y) = \begin{cases} 0.5y^2, & |y| \le 1 \\ |y| - 0.5, & otherwise \end{cases} \quad (10)$$

Where, $\hat{s}_i$ denotes the label of whether the video frame is an anomaly or not; s_i is the anomaly score predicted by the network and y represents the difference between s_i and $\hat{s}_i$.

4 EXPERIMENTAL RESULTS AND ANALYSES

In this paper, the effectiveness of the proposed algorithm is tested on three datasets, CUHK Avenue, USCDPed2 and LAD2000, and validated by comparison with eight methods and ablation experiments.

4.1 Experimental Data

LAD2000 dataset [8] Contains 2000 video sequences divided into 1440 training sets and 560 test sets. The resolution is 226×400 and the frame rate is 25fps. includes video sequences collected from public websites including YouTube, Youku and Tencent videos and existing motion recognition databases, as well as some normal activities and abnormal behaviours in plazas and schools recorded with digital cameras. UCSD Ped2 dataset [19] contains 28 video sequences divided into 14 training sets and 14 test sets. The resolution is 238×158 and the frame rate is 10fps, totalling 4950 video frames.The UCSD Ped2 dataset is characterised by the inclusion of complex scenarios such as occlusions and crossings between pedestrians. CUHK Avenue dataset [20]contains 37 video sequences divided into 16 training sets and 21 test sets. Each sequence has a resolution of 640×360 and a frame rate of 25fps, totalling 30,652 video frames, each with pixel-level annotations. The video sequences were shot on city streets and contain a variety of different situations, such as pedestrians crossing the street, vehicles travelling, etc.

4.2 Experimental Setup

The CPU for this experiment is Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz*12, GPU is NVIDIA Tesla V100-SXM2 -32GB with 48GB RAM, and the software environment is PyTorch 1.2.0, Python 3.8 and CUDA 10.0. In this paper, the Area Under Curve (AUC) of the Receiver operating characteristic (ROC) curve at frame level is used as an evaluation metric, and the experimental results are calculated by comparing the frame level detection results with the real labels at frame level. Using the video classification dataset Kinetics-400 [21] Pre-train the I3D backbone network. Using the Adam optimiser [22] The model parameters are updated with the learning rate set to 3e-4, weight decay set to 5e-4 and batch size set to 60. the video sequence is divided into 16 non-overlapping frame segments fed into the I3D backbone network to extract features on the RGB and optical flow inputs. The threshold for binarisation of the anomaly scores is set to 0.5, the number of model iterations is 12000, the batch size is 200 and the learning rate is 1e-4. The SENet module reshapes the fused features to 5×60×1024 and feeds them into ConvLSTM.The hidden layer channels of ConvLSTM are set to 128, the convolution kernel is 3×3, and the step size is 1×1.The dimensionality of the output features in ConvLSTM is 4×4×128.The obtained global spatio-temporal contextual features are reshaped to 1024-dimensional vectors and then sent to the three fully convolutional layers to predict the final anomaly score, where the dimensions of the first two full convolutional layers are set to 1024 and 512, respectively. the input video parameters are set to m=16 and N=5.

4.3 Ablation Experiments

In this paper, the feature fusion method used, SENet, is compared with several other feature fusion methods. The results of the ablation experiments are shown in Table 1, from which it can be seen that the attentional feature fusion method Concat+SENet used in this paper outperforms other fusion methods. The accuracy on the CUHK Avenue dataset (86.44%) is improved by 1.03% compared to the UCSD Ped2 dataset, which is almost the same as the accuracy on the UCSD Ped2 dataset, and the accuracy on the LAD2000 dataset (85.25%) is improved by 1.24% compared to the UCSD Ped2 dataset, which is due to the fact that the Concat method is only through the simple feature splicing and lacks of attention on the correlation among different features, and cannot reflect the important features through the weight values. This is due to the fact that the Concat method lacks attention to the correlation between different features through simple feature splicing and does not reflect the selection of important features through the weight values.

Table 1 Ablation Experiment Results

Fusion Methods	AUC (%)			Param(M)	FLOPs (M)
	CUHK Avenue	UCSD Ped2	LAD2000		
Concat[8]	86.44	93.83	85.25	-	-
Concat+MS-CAM[13]	56.64	51.57	46.58	$4 \times D \times d + d^2$	17.29
Concat+MHSA[23]	86.41	94.70	85.93	$D \times d + 4 \times d^2$	94.90
Concat+SENet[22]	**87.47**	**93.72**	**86.49**	**2 x D x d**	**1.03**

Concat+SENet improves the accuracy of Concat+MHSA method by 1.06% compared to Concat+MHSA method on the CUHK Avenue dataset (86.41%), which is nearly the same as that on the UCSD Ped2 dataset (94.70%), and improves the accuracy by 0.54% compared to that on the LAD2000 dataset (85.93%). And the number of parameters of LAFF is much reduced compared to the Concat+MHSA method. The FLOPs of the feature fusion method used in this paper (1.03M) are reduced by 93.87% compared to the Concat+MHSA method (94.90M). From the comparison of Concat and Concat+MHSA feature fusion by adding attention has better detection accuracy compared to simple linear splicing on some datasets, and Concat+SENet improves the feature fusion while ensuring a lower number of parameters and FLOPs.

4.4 Comparative Experiments

In order to prove the effectiveness of the proposed method in the field of video anomaly detection, the algorithm of this paper is compared with seven existing methods and the comparison results are shown in Table 2. The unsupervised anomaly detection methods include AST-AE [2], Chang et al. [3], MNAD [24], weakly supervised detection methods including DeepMIL [5], AR-Net [6] and MLEP [7], fully supervised detection methods including Wan et al. [8].

Table 2 Comparison Results with Existing Methods

Methods	CUHK Avenue	UCSD Ped2	LAD2000
AST-AE[2]	85.20	96.60	-
Chang et al.[3]	87.10	96.70	-
DeepMIL[5]	87.53	90.19	70.18
AR-Net [6]	89.31	93.64	79.84
MLEP[7]	89.20	-	50.57
Wan et al.[8]	89.33	95.12	86.28
MNAD[24]	82.80	90.20	45.84
Our method	87.47	93.72	86.49

The detection accuracy of this paper's algorithm on the CUHK Avenue dataset (87.47%), compared to unsupervised AST-AE et al. (85.20%) is improved by 2.27%, and compared to the weakly supervised DeepMIL (87.53%) is nearly the same. The detection accuracy on the UCSD Ped2 dataset (93.72%), compared to the weakly supervised DeepMIL (90.19%) improved by 3.53%. The detection accuracy obtained on the LAD2000 dataset (86.49%), compared to the unsupervised MNAD (45.84%) improved by 40.65%, compared to the weakly supervised AR-Net (79.84%) by 6.65%, compared to the fully supervised Wan et al. (86.28%) by 0.21%. The above comparison results can illustrate the effectiveness of the algorithm in this paper.

From the above comparison of the results of each dataset, this paper's algorithm uses fully supervised learning, compared to weakly supervised methods using video-level labels to train the model and unsupervised methods to learn anomalies by features, there are certain advantages, and also overall illustrates the effectiveness of this paper's algorithm on the basis of the previous validation of SENet and ConvLSTM.

5 CONCLUSION

In this paper, we propose a video anomalous behaviour detection method based on attentional feature fusion, for the problems of insufficient cross-domain fusion capability and poor spatio-temporal perception of a single convolution, the cross-domain fusion mechanism constructed by the SENet module is used to compensate for the insufficient feature fusion capability of the current method, and the effect of spatio-temporal perception after feature fusion is achieved through the ConvLSTM module, so as to improve the performance of the video anomalous detection effectively. The experimental results on the three datasets show that the proposed method is significantly better than the existing methods in terms of accuracy, and significantly reduces the number of parameters compared to the previous models, achieving a balance between performance and cost. The correlation between features will be further explored in the next research work to improve the detection performance of abnormal behaviours.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported in part by the Social Science Fund of Guangxi under Grant 23FTQ005.

REFERENCES

[1] Jiancong Wang. Deep learning based multi-module joint video anomaly detection research. Guilin University of Electronic Science and Technology, 2023.
[2] Liu Yang, Yang Dingkang, Wang Yan, et al. Generalised video anomaly event detection: systematic taxonomy and comparison of deep models. arXiv preprint arXiv: 2302.05087, 2023.
[3] Chang Yunpeng, Tu Zhigang, Xie Wei, et al. Video anomaly detection with spatio-temporal dissociation. Pattern Recognition, 2022, 122: 108213.

[4] Tudor Ionescu r, Smeureanu Sorina, Alexe Bogdan, et al. Unmasking the abnormal events in video. Proceedings of the 2017 IEEE international conference on computer vision. Piscataway: IEEE, 2017: 28952903.
[5] Sultani Waqas, Chen Chen, Shah Mubarak. Real-world anomaly detection in surveillance videos. Proceedings of the 2018 IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2018: 6479-6488.
[6] Wan B, Fang Y, Xia X, et al. Weakly supervised video anomaly detection via centre-guided discriminative learning//2020 IEEE international conference on multimedia and expo (ICME). IEEE, 2020: 1-6.
[7] Liu Wen, Luo Weixin, Li Zhengxin, et al. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies. Proceedings of the 2019 International Joint Conferences on Artificial Intelligence. Freiburg: IJCAI, 2019: 3023-3030.
[8] Wan Boyang, Jiang Wenhui, Fang Yuming, et al. Anomaly detection in video sequences: a benchmark and computational model. IET Image Processing, 2021, 15(14): 3454-3465.
[9] Zhou Jiapeng. Research on Unsupervised Learning-based Domain-adaptive Semantic Segmentation Methods. China University of Mining and Technology, 2023.
[10] Zou Wei, Zhang Dong, Lee Dahjye. A new multi-feature fusion based convolutional neural network for facial expression recognition. Applied Intelligence, 2022, 52(3): 2918-2929.
[11] Cheng Xianggui, Liu Zhao, Guo Fang. Video Abnormal Event Detection Combining Dual-stream I3D and Attention Mechanism. Information and Computer (Theoretical Edition), 2022, 34(24): 65-68.
[12] Dai Yimian, Gieseke Fabian, Oehmcke Stefan, et al. Attentional feature fusion. Proceedings of the 2021 IEEE/CVF winter conference on applications of computer vision. Piscataway: IEEE, 2021: 3560-3569.
[13] Huang Shaonian, WEN Peiran, QUAN Qi, et al. Lightweight video anomaly detection based on multi-branch aggregation frame prediction. Journal of Graphics, 2023, 44(6): 1173.
[14] Le Viettuan, Kim Yongguk. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence, 2023, 53: 3240-3254.
[15] Ye Wenbing, Zhan Shihua. Anomaly Detection Method of Network Traffic Based on MHA-BiLSTM. Modern Information Technology, 2024, 8(02): 65-69.
[16] Liu Wei, Rabinovich Andrew, C.Berg Alexander. Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579. 2015.
[17] Shi Xingjia, Chen Zhourong, Wang Hao, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Proceedings of the 2015 Conference on Neural Information Processing Systems. San Diego: NeurIPS, 2015.
[18] Szegedy Christian, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions. Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston. Washington: IEEE Computer Society, 2015: 1-9.
[19] Yan Shanwu, Xiao Hongbing, Wang Yu, et al. Video anomaly detection by fusing pedestrian spatiotemporal information. Journal of Graphics, 2023, 44(1): 95.
[20] Roka Sanjay, Diwakar Manoj: a deep convolutional encoder-decoder architecture for abnormality detection in video surveillance. Cluster Computing, 2024: 1-16.
[21] Li Kunchang, Wang Yali, He Yinan, et al. Uniformerv2: Unlocking the potential of image vits for video understanding. Proceedings of the 203 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 1632-1643.
[22] Hu Jie, Shen Li, Sun Gang. Squeeze-and-excitation networks. Proceedings of the 2018 IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2018: 7132- 7141.
[23] Lv Hui, Yue Zhongqi, Sun Qianru, et al. Unbiased multiple instance learning for weakly supervised video anomaly detection. Proceedings of the 2023 IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE, 2023: 8022-8031.
[24] Park Hyunjong, Noh Jongyoun, Ham Bumsub. Learning memory-guided normality for anomaly detection. Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE, 2020: 14372-14381.