Intelligent Automation & Soft Computing
DOI: 10.32604/iasc.2022.029535
Article

Tech Science Press

Criss-Cross Attention Based Auto Encoder for Video Anomaly Event Detection

Jiaqi Wang[1], Jie Zhang[2], Genlin Ji[2,*] and Bo Sheng[3]

[1]School of Mathematical Sciences, Nanjing Normal University, Nanjing, 210023, China
[2]School of Computer and Electronic Information, Nanjing Normal University, Nanjing, 210023, China
[3]Department of Computer Science, University of Massachusetts Boston, Boston, 02125, USA
*Corresponding Author: Genlin Ji. Email: glji@njnu.edu.cn
Received: 05 March 2022; Accepted: 12 April 2022

Abstract: The surveillance applications generate enormous video data and present challenges to video analysis for huge human labor cost. Reconstruction-based convolutional autoencoders have achieved great success in video anomaly detection for their ability of automatically detecting abnormal event. The approaches learn normal patterns only with the normal data in an unsupervised way due to the difficulty of collecting anomaly samples and obtaining anomaly annotations. But convolutional autoencoders have limitations in global feature extraction for the local receptive field of convolutional kernels. What is more, 2-dimensional convolution lacks the capability of capturing temporal information while videos change over time. In this paper, we propose a method established on Criss-Cross attention based AutoEncoder (CCAE) for capturing global visual features of sequential video frames. The method utilizes Criss-Cross attention based encoder to extract global appearance features. Another Criss-Cross attention module is embedded into bi-directional convolutional long short-term memory in hidden layer to explore global temporal features between frames. A decoder is executed to fuse global appearance and temporal features and reconstruct the frames. We perform extensive experiments on two public datasets UCSD Ped2 and CUHK Avenue. The experimental results demonstrate that CCAE achieves superior detection accuracy compared with other video anomaly detection approaches.

Keywords: Video anomaly detection; bi-directional long short-term memory; convolutional autoencoder; Criss-Cross attention module

1 Introduction

The ubiquitous surveillance cameras in public areas such as streets, banks and malls have produced massive amount of video data. It is time consuming and infeasible for human observers analyzing and monitoring every video stream. Video Anomaly Detection (VAD) is important in intelligent surveillance systems which can automatically detect appearance and motion anomaly of objects that deviate significantly from the normality [1]. VAD aims to associate each frame with an anomaly score for the temporal variation, a spatial score to localize the anomaly in space, and identify the type of anomaly [2]. Although VAD has been studied for several decades, the task still remains challenging. (1) The anomaly

events are much less than normal ones in real-world scenarios [3]. (2) The definition of abnormal samples is not clear. For example, crowd gathering is normal in supermarket while abnormal in the context of epidemic prevention. (3) Anomalous behaviors and items are diverse and the distribution of anomaly patterns is unknown in advance [4]. It is very difficult to detect all possible anomalous samples. The imbalance, ambiguous and diversity characteristics of video data make it impractical to gather labeled data of all types of possible anomalies. To address this problem, VAD is always treated as an unsupervised task by exploiting the regular patterns only with the normal data.

Reconstruction-based models are the most common strategies in unsupervised learning. The basic idea is to reconstruct normal data with low reconstruction error in the training phrase. During testing, the distinctive encoded patterns are detected as anomalies. Early anomaly detection methods are mainly relied on handcrafted feature engineering with machine learning techniques [5]. In recent years, deep learning-based reconstruction models are popular for the capability of extracting features and training models in a unified framework. Videos are high dimensional signals with both spatial structure and temporal variations [2]. For unsupervised representation learning, a variety of models have been proposed to combine Convolutional Neural Network (CNN) and autoencoder to automatically model both the appearance and motion features. However, these methods have limitations in global information extraction, because CNN concentrates on local information and will lead to information loss on remote features. In addition, 2-dementional CNN is suitable for images, but it is incapable to capture the temporal information for consecutive video frames.

Surveillance videos change over time. The chain-like building blocks of Long Short-Term Memory (LSTM) with forget, input and output gates can regulate long-term sequence pattern recognition [6]. The variant of convolutional LSTM (ConvLSTM) models have been proposed with autoencoder to reconstruct and predict temporal features for sequential video frames [7]. Although deep convolutions are strong in visual features learning and LSTM can capture temporal information, the ConvLSTM models are limited in local receptive fields of CNN which only provide a few surrounding contextual information.

To better capture long-range global dependencies, many researches optimize the encoder and decoder through an attention mechanism [8]. Self-attention [9] is an efficient mechanism which allows each query position in the input sequence to perceive all positions and take their weighted average. The model implies the importance to the effective areas and suppresses other irrelevant areas. In this way self-attention module can obtain full-image contextual information. But the generation of attention maps leads to a very high computational complexity, because we have to measure the relationships for each pair of pixels. Recently, Criss-Cross attention [10] is put forward as an efficient way to harvest full-image contextual information for all the pixels only on its criss-cross path.

To handle the issues mentioned above, we propose a novel method which combines Criss-Cross attention module with bi-directional ConvLSTM and autoencoder for video anomaly detection. The method utilizes an encoder with Criss-Cross module to extract global appearance features. Global temporal features are then calculated by adding a Criss-Cross based bi-directional ConvLSTM network between the encoder and decoder. A decoder is used to fuse the two kinds of features and reconstruct frames for the goal of capturing global features in spatial and temporal space simultaneously. Given several consecutive frames, the method learns the normal patterns well, while the anomaly frame is expected to be distorted and blurry.

The main contributions of this paper are summarized as follows:

1. We propose a novel CCAE method for video anomaly detection. By combining Criss-Cross attention module with bi-directional ConvLSTM and autoencoder, the method is able to learn global appearance features and global temporal features.

2. A smooth L1 loss is utilized as intensity loss to compare two video frames pixel by pixel for reconstruction. Smooth L1 loss is smoother near zero than L1 loss, and can prevent from exploding gradient compare to L2 loss.
3. CCAE can accurately detect abnormal events with frame level AUC of 95.3% for UCSD Ped2 dataset and 84.0% for CUHK Avenue dataset. The experimental results show that the proposed method attains competitive detection accuracies compared with other VAD methods.

The rest of the paper is organized as follows. Section 2 discusses the brief review of related works on video anomaly detection and attention mechanism. Section 3 describes the details of the proposed CCAE method. The evaluation of experimental results and comparison with existing methods are presented in Section 4. At the end of this paper we conclude our work and discuss the future research directions.

2 Related Works

2.1 Video Anomaly Detection

Due to the scarce of anomaly data and difficulties of annotations, video anomaly detection has been formulated mainly in unsupervised settings. Previous anomaly detection methods are mainly based on handcrafted feature engineering. Kim et al. [11] capture the distribution of local optical flow patterns with a Mixture of Probabilistic Principal Component Analyzers (MPPCA) [12], then use Markov Random Field (MRF) graph to detect abnormal patterns in incoming video clips. Mahadevan et al. [13] model normal crowd behavior for each spatial-temporal block using Mixtures of Dynamic Textures (MDT) [14], then temporal anomalies are equated to events of low-probability, while spatial anomalies are handled using discriminant saliency. Mehran et al. [15] detect and localize abnormal crowd behavior using the social force (SF) model. Adam et al. [16] detect abnormal events by multiple monitors which utilize histograms to measure the probability of optical flow in a local patch. The handcrafted features can accurately model both spatial and temporal information. However, they are task-specific and require prior knowledge for feature designing. Therefore, the handcrafted features are difficult to adapt to other scenarios [17] and are impractical in real scenarios.

Nowadays, deep learning models have been shown to perform well in video anomaly detection tasks. For example, CNN can effectively extract high-level features by local kernels. Sabokrou et al. [18] use Fully Convolutional Neural Network (FCNN) for detecting and localizing anomalies. Some researchers also train CNN on large-scale ImageNet to obtain the feature representation in video object tracking to solve the problem of insufficient training data [19]. Recently, many unsupervised video anomaly detection models are proposed based on deep autoencoder architecture, which is composed of an encoder to compress the input vector into a low-dimension embedding, and a decoder to reconstruct the dense vector back to the input vector [20]. Hasan [21] build a 2-dimentional Convolutional AutoEncoder (ConvAE) to model the normal videos by stacking the frames. U-net [22] is and end-to-end encoder-decoder architecture for biomedical image segmentation. Sabokrou et al. [23] introduce a novel cubic-patch-based model based on autoencoder to reconstruct input video patch. But 2-dimentional convolutional operation fails to capture temporal cues of video frames. Although Deepak et al. [24] try to capture information on the temporal dimension by using 3-dimentional kernels, it is still inadequate to accurately detect the sequential anomalous samples. Modeling temporal patterns in a timely manner has remained challenging.

Learning temporal features in VAD has attracted many researchers. Inspired by the temporal capability of LSTM for sequential video frames, some works combine ConvAE and LSTM to model spatial and temporal normal patterns simultaneously. Chong et al. [25] use a stack of ConvAE to capture spatial structures, and then the video representation is fed into a stack of ConvLSTM in autoencoder architecture for temporal patterns. Luo et al. [26] design a parse coding inspired deep recurrent neural network

autoencoder framework for alleviating the hyper-parameters selection and dictionary training in temporally-coherent sparse coding. While effective, ConvLSTM is constrained by the size of convolution kernels of CNN. The models focus on the local information and cannot fuse the remote features.

2.2 Attention Mechanism

Attention module has achieved great success in many computer vision tasks to aggregate global contextual information for better feature representation. The methods can be classified as channel-wise attention [27] and spatial-wise attention [28]. Hu et al. [27] propose a Squeeze-and-Extraction (SE) block to calculate the channel-wise feature maps. Gong et al. [28] propose a Memory-augmented AutoEncoder (MemAE) to encode the latent vector as query to obtain the soft addressing weights. Sun et al. [29] propose a multi-feature learning model with global feature and enhanced local attention for vehicle re-identification in video surveillance. We have known that convolutions and recurrent operations process the local neighborhood at a time, either spatially or temporally; then the long-term dependencies are modeled by repeatedly applying the local operations. In contrast to these ineffective local models, Wang [30] propose a non-local module to capture long-range dependencies by computing interactions between any two positions. The model guarantees that a pixel at any position can perceive contextual information of all other pixels. The non-local modules can be combined with other models easily. However, the non-local methods are always with huge attention maps and thus are computational complexity. To address this problem, a more efficient Criss-Cross attention [10] is proposed to aggregate the contextual information of each pixel only in its horizontal and vertical directions. By stacking two consecutive Criss-Cross modules, each position of input can collect contextual information from all other pixels.

3 Method

In this section, we employ the proposed CCAE to learn the normal patterns of normal videos. The diagram of video anomaly detection process is illustrated in Fig. 1. A sequence of frames f_{t-3}, f_{t-2}, f_{t-1}, f_t are given, we adopt CCAE to reconstruct the last frame f_t as f_{rec}. In the training stage, CCAE is trained by minimizing the loss function calculated between the reconstructed frame f_{rec} and the real frame f_t. In this way the normal patterns are learned from normal training samples. During testing, we calculate the regularity score of testing frame f_t based on the reconstructed frame f_{rec} and the ground truth f_t. Testing frame f_t will be regarded as abnormal if it deviates significantly from the normal pattern.

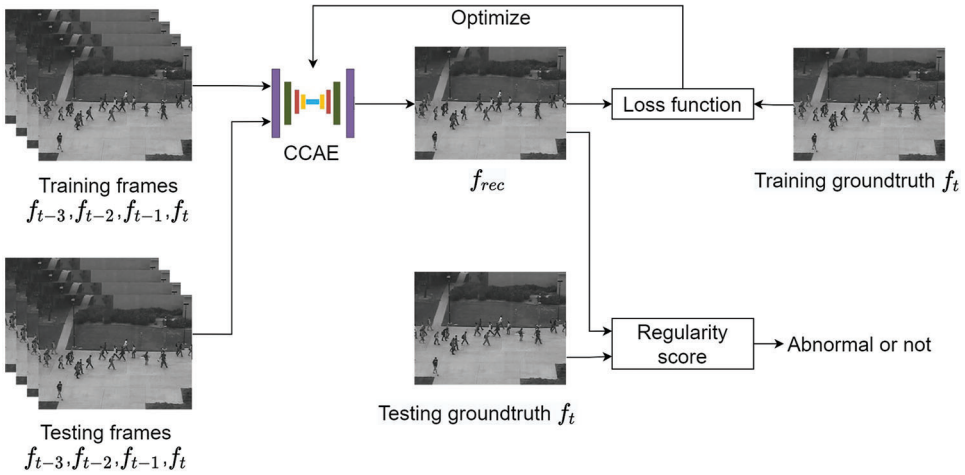

Figure 1: The diagram of video anomaly detection process

3.1 Criss-Cross Attention Based Auto Encoder (CCAE)

The proposed CCAE is shown in Fig. 2. The method is composed of Criss-Cross attention encoder, Criss-Cross attention based Bi-directional ConvLSTM (CCBiLSTM) and a decoder. The Criss-Cross attention encoder extracts continuous appearance features with an encoder and obtains the global appearance features with Criss-Cross attention module; then the continuous appearance features are fed into the CCBiLSTM to capture global temporal features; the extracted global appearance features and global temporal features are fused together and used to reconstruct the frames. Inspired by U-net [22], each layer of the encoder-decoder in CCAE is added with a skip connection to obtain the features of the same layer, which can retain more context semantic information.

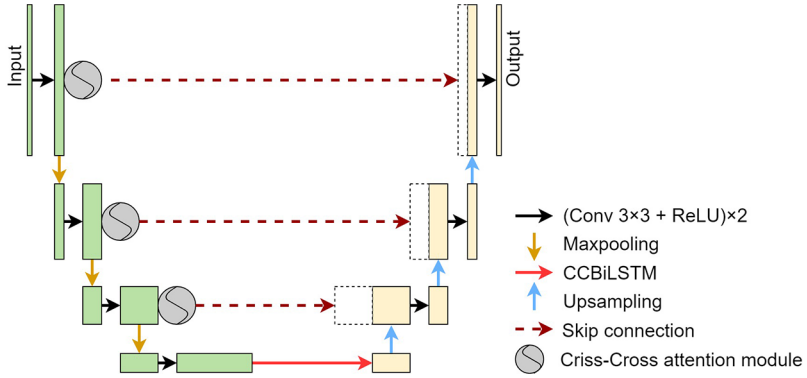

Figure 2: The overview of CCAE

3.1.1 The Criss-Cross Attention Module

We first introduce the Criss-Cross attention module [10] which is utilized to aggregate global contextual information of video appearance and temporal features in CCAE. The module can capture the contextual information of each pixel in its horizontal and vertical directions which is very effective. For example, if an input feature map H is fed into the Criss-Cross module, a new feature map H' will be generated by collecting the contextual information for each pixel in its criss-cross path.

The details of Criss-Cross Attention module are shown in Fig. 3. Given a local feature map H, the module first applies two convolutional layers with 1×1 filters on $H \in R^{C \times H \times W}$ to generate two feature maps $Q \in R^{C' \times H \times W}$ and $K \in R^{C' \times H \times W}$. We further generate an attention map $A \in R^{(H+W-1) \times H \times W}$ via Affinity operation. At each position u in the spatial dimension of Q, we can obtain a vector $Q_u \in R^{C'}$. Meanwhile, we can also obtain the set $\Omega_u \in R^{(H+W-1) \times C'}$ by extracting feature vectors from K which are in the same row or column with position u. $\Omega_{i,u} \in R^{C'}$ is the i_{th} element of Ω_u. The Affinity operation is then defined as follow:

$$d_{i,u} = Q_u \Omega_{i,u}^T$$

where $d_{i,u}$ is the degree of correlation between feature Q_u and $\Omega_{i,u}$. Then a softmax layer is applied over the channel dimension to calculate the attention map A.

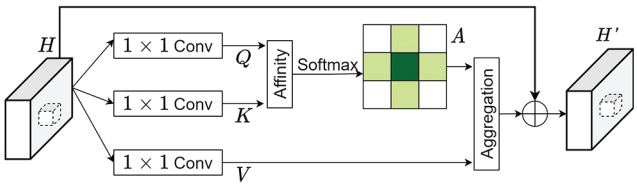

Figure 3: The details of Criss-Cross attention module

The third convolutional layer with 1×1 filters is applied on H to generate $V \in R^{C\times H\times W}$ for feature adaptation. We can obtain a vector $V_u \in R^C$ and a set $\Phi_u \in R^{(H+W-1)\times H\times W}$. The set Φ_u is a collection of feature vectors in V which are in the same row or column with position u. The contextual information is collected by an Aggregation operation defined as follow:

$$H_u^{'} = \sum_{i\in|\Phi_u|} A_{i,u}\Phi_{i,u} + H_u \tag{2}$$

where $H_u^{'}$ is a feature vector in $H' \in R^{C\times H\times W}$ at position u and $A_{i,u}$ is a scalar value at channel i and position u in A. The contextual information is added to local feature H to augment the pixel-wise representation. Therefore, it has a wide contextual view and selectively aggregates contexts according to the spatial attention map. Each position in the feature map is sparsely connected with others which are in the same row and the same column in Criss-Cross attention module, leading to the global attention map only has about $2\sqrt{N}$ weights rather than N in non-local module. By stacking two consecutive Criss-Cross models, each position can perceive the full contextual information from all the pixels of the input. In this way, CCAE can effectively capture long-range global dependencies of video features through the Criss-Cross attention mechanism.

3.1.2 Criss-Cross Attention Based Auto Encoder

In CCAE, the Criss-Cross attention encoder extracts the appearance features of the video frames by convolution operation, and the size of feature map is reduced to 1/2 of the original size by maxpooling. With the number of channels remaining consistent, totally four iterations are performed to obtain the appearance features in different scales. Then, these appearance features are fed into Criss-Cross attention module to obtain the global appearance features in different scales. The fourth extraction of the appearance features is sent to CCBiLSTM to extract the global temporal features. In the final step, a decoder fuses global appearance features and global temporal features by deconvolution and connection, and generates high-quality reconstructed frames.

The details of CCBiLSTM is shown in Fig. 4. The Criss-Cross attention module not only captures the global spatial relationship of a single video frame, but also captures the global temporal dependency between consecutive video frames. In this way, attention can be allocated in the whole spatial region and time dimension to improve the utilization of temporal and appearance features. The continuous appearance features are input into bi-directional ConvLSTM to extract the temporal features between video frames, and then the global temporal features are obtained by weighting the temporal features through the Criss-Cross attention module. The bi-directional ConvLSTM network consists of forward ConvLSTM and reverse ConvLSTM, which generate forward and reverse feature vectors, and these two vectors can be connected to generate video frame feature. Then, the global temporal features are obtained by weighting the features with Criss-Cross attention modules, as follows:

$$\overrightarrow{h_t} = ConvLSTM^f([x_1, x_2, x_3, \ldots x_t]) \tag{3}$$

$$\overleftarrow{\mathrm{h_t}} = \mathrm{ConvLSTM^b}([\mathrm{x_1}, \mathrm{x_2}, \mathrm{x_3}, \ldots \mathrm{x_t}]) \tag{4}$$

$$\mathrm{h_t} = \mathrm{cat}\left[\overrightarrow{\mathrm{h_t}}, \overleftarrow{\mathrm{h_t}}\right] \tag{5}$$

$$\mathrm{h_{global\ t}} = \mathrm{CCatt(h_t)} \tag{6}$$

where $\overrightarrow{h_t}$ represents the temporal features obtained by forward ConvLSTM $ConvLSTM^f$, $\overleftarrow{h_t}$ denotes the temporal features obtained by reverse ConvLSTM $ConvLSTM^b$. $[x_1, x_2, x_3, \ldots x_t]$ represents the video frame features of consecutive t frames extracted by the encoder, h_t is the time sequence feature which

concatenates $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$. $h_{global\ t}$ represents the global temporal features obtained after Criss-Cross attention operation.

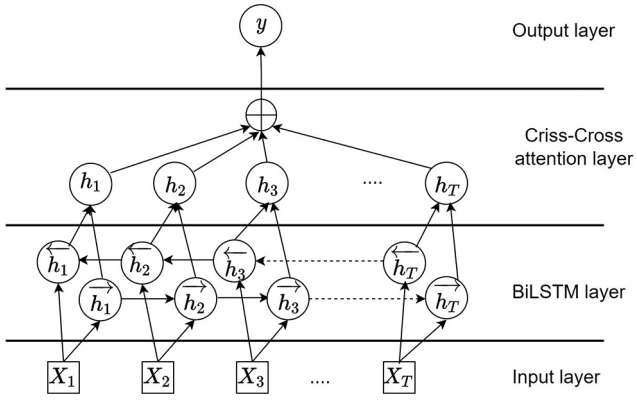

Figure 4: The details of CCBiLSTM

3.2 Loss Function

Intensity loss and gradient loss are employed into loss function of CCAE so that the reconstructed video frames can be closer to the real frames. The intensity loss L_{int} uses smooth L1 loss [31] to compare two video frames pixel by pixel. Compared with L1 loss, smooth L1 loss is robust to outliers and is smoother near zero. What is more, the smooth L1 loss can prevent from exploding gradient of L2 loss [32]. We define x as the distance between the reconstructed frame pixel $\hat{I}_{i,j}$ and real frame pixel $I_{i,j}$, then the loss function is calculated pixel by pixel as follow:

$$L_{int} = \begin{cases} \frac{1}{2}x^2 & if\ |x| < 1 \\ |x| - \frac{1}{2} & otherwise \end{cases} \tag{7}$$

The purpose of gradient loss L_{gd} is to ensure that the gradients of the reconstructed frame and the real frame are close. The gradient loss calculates the distance between the pixels in the reconstructed frame and the adjacent pixels in the real frame, which is shown as follow:

$$L_{gd} = \sum_{i,j}\left(\left\|\left|\hat{I}_{i,j} - \hat{I}_{i-1,j}\right| - \left|I_{i,j} - I_{i-1,j}\right|\right\|_1 + \left\|\left|\hat{I}_{i,j} - \hat{I}_{i,j-1}\right| - \left|I_{i,j} - I_{i,j-1}\right|\right\|_1\right) \tag{8}$$

where i and j represent the horizontal and vertical coordinates of the pixels in the video frame.

In the training stage, we combine intensity constraint and gradient constraint to improve the reconstruction ability of CCAE by minimizing the loss function L_{re}. The combined loss function is as follow:

$$L_{re} = \lambda_{int}L_{int} + \lambda_{gd}L_{gd} \tag{9}$$

where λ_{int} and λ_{gd} are used for weighted loss.

3.3 Regularity Score

For abnormal frames, there will be more reconstruction errors compared with the ground truth. In other words, the probability of a frame to be abnormal impacts the quality of the reconstructed frame. It is intuitive

to calculate the anomaly score by measuring image quality. We use the Peak Signal Noise Ratio (PSNR) for image quality assessment.

$$PSNR(I,\hat{I}) = 10log_{10}\frac{(max_{\hat{I}})^2}{\frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N}(I_{i,j}-\hat{I}_{i,j})^2} \tag{10}$$

where $I_{i,j}$ is the ground truth of $pixel(i,j)$, and $\hat{I}_{i,j}$ is the reconstructed $pixel(i,j)$. M and N are the height and width of the frame.

A higher PSNR value indicates a higher quality of the frame, which means it is more likely to be normal. We normalize the PSNR value into $[0, 1]$ for frame t and the regularity score is calculated as follows:

$$Score(t) = \frac{PSNR_t - min\ PSNR}{max\ PSNR - min\ PSNR} \tag{11}$$

where $PSNR_t$ is its original PSNR value. *max PSNR* and *min PSNR* are the maximum and minimum PSNR values in the current video. We predict a frame to be normal or abnormal based on its regularity score.

4 Experiments

4.1 Experiment Settings

To verify the effectiveness of CCAE, we implement the experiments in Pytorch with four NVIDIA GPUs. Input frames are resized to the resolution of 256*256. In the training stage, the learning rate is set to 0.0003 and batch size is set to 8. The number of training epochs is set to 200. The weighted coefficients in the training loss are set as $\lambda_{int} = 1$ and $\lambda_{gd} = 0.1$.

The experiments are conducted on two public video anomaly detection datasets, UCSD Ped2 and CUHK Avenue. The anomalies refer to objects and events that do not conform to expectations. The UCSD Ped2 dataset contains 16 training videos and 12 testing videos. The videos are captured with a stationary camera. The dataset contains people walking normally in a pedestrian sidewalk. The anomalies are about appearance of non-pedestrian entities, such as riding a bike and driving a car in pedestrian area. The CUHK Avenue dataset consists of 16 training videos with normal activities and 21 testing videos. Anomalous events are related to people running, walking in the wrong direction and throwing objects. The details of the two datasets are shown in Tab. 1.

Table 1: Details of two public video anomaly detection datasets

Datasets	Scenarios	Anomalies	Resolution
UCSD Ped2	Sidewalk	Appearance of non-pedestrian entities and anomalous pedestrian behaviors	360 × 240
CUHK Avenue	Campus	Strange action, wrong direction and abnormal object	640 × 360

4.2 Evaluation Metrics

The area under ROC curve (AUC) is adopted as metric to evaluate the detection accuracy of CCAE. ROC curve is obtained by varying the threshold of the anomaly score for each frame-wise reconstruction. A higher AUC value indicates that the detection performance is better. Besides, the equal error rate (EER) is also reported as the percentage of misclassified frames.

4.3 Experimental Results

The examples of reconstructed frames and the actual frames are visualized in Fig. 5. It is shown that CCAE has good reconstruction effect in the normal scenes, while in abnormal scenes the reconstructed frames are blurry and distorted. The abnormal areas are circled in red. In this way, CCAE can judge whether the frame is abnormal or not.

Figure 5: The comparison between reconstructed and actual frames

As shown in Figs. 6 and 7, the regularity score is calculated to visualize the performance on CUHK Avenue and UCSD Ped2 datasets. The positions of normal and abnormal frames can be seen directly. This means that the smaller the regularity score is, the higher the probability of abnormal frame is. The red area in the graph represents anomalies when the exception occurs in the dataset.

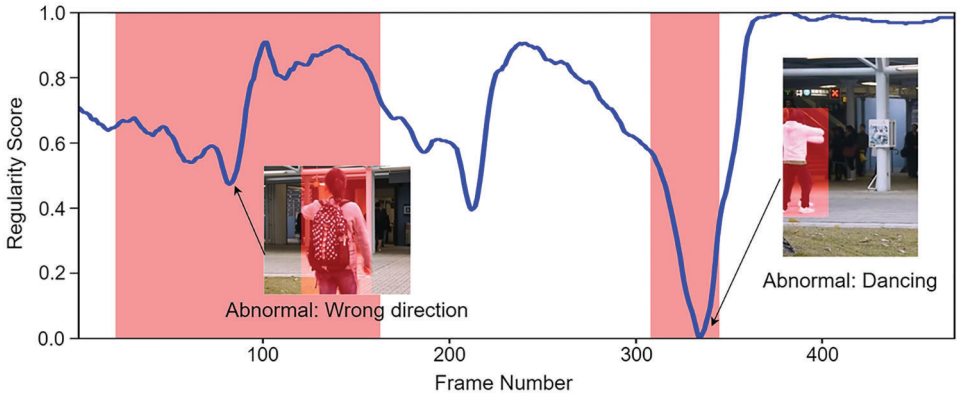

Figure 6: Regularity score visualization for CUHK Avenue dataset

Tab. 2 shows the AUC and EER of CCAE compared with other video anomaly detection methods. For UCSD Ped2 dataset, we can see that deep learning based methods achieve higher AUC and lower EER than handcraft based methods such as SF and MPPCA. The results verify the excellent learning ability of deep models. Specifically, ConvAE and STAE perform better than handcraft based models, which implies that the models can benefit from integrating deep CNN, autoencoder and convLSTM. Compared with ConvAE and STAE, GMMAE is able to combine Gaussian mixture model and variational autoencoder. The knowledge fusion helps to improve the anomaly detection accuracy. As expected, CCAE attains competitive detection accuracy compared with other deep learning based methods. It is clearly that utilizing only convolutional autoencoder and LSTM to learn spatial and temporal features is not adequate.

The experimental results show that the extraction of global appearance features and global temporal features based on Criss-Cross attention module can effectively improve the anomaly detection performance.

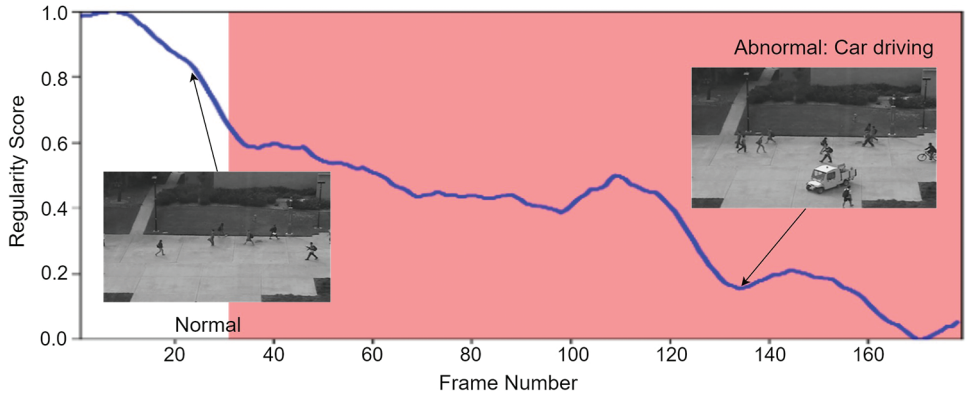

Figure 7: Regularity score visualization for UCSD Ped2 dataset

Table 2: Comparison with other anomaly detection methods

Methods	UCSD Ped2		CUHK Avenue	
	AUC(%)	EER(%)	AUC(%)	EER(%)
MPPCA [31]	69.3	30.0	–	–
SF [15]	55.6	42.0	–	
SF+MPPCA [11]	61.3	36.0	–	–
ConvAE [21]	90.0	21.7	70.2	25.1
STAE [25]	91.2	16.7	80.9	24.4
GMMAE [17]	92.2	12.6	83.4	22.7
MemAE [29]	94.1	–	83.3	–
TSC [26]	92.21		83.48	
Deep STAE [24]			83	
CCAE	**95.3**	**10.8**	**84.0**	**21.9**

4.4 Ablation Experiment

In order to analyze the effectiveness of Criss-Cross attention module, we compare CCAE with bi-directional long short term memory autoencoder (BiAE). Unlike CCAE, BiAE only inputs local features extracted by encoder and bi-directional long short term memory network directly into decoder for reconstruction. The AUC and EER are listed in Tab. 3. We can see that Criss-Cross attention module can achieve a positive effect on anomaly detection. Fig. 8 shows feature heatmap made by CCAE and BiAE on the two datasets. We can see that capturing long-range global dependencies with Criss-Cross attention module can enhance attention in key areas.

Table 3: Comparison of CCAE with BiAE

Methods	UCSD Ped2		CUHK Avenue	
	AUC(%)	EER(%)	AUC(%)	EER(%)
BiAE	92.3	15.4	80.8	23.4
CCAE	**95.3**	**10.8**	**84.0**	**21.9**

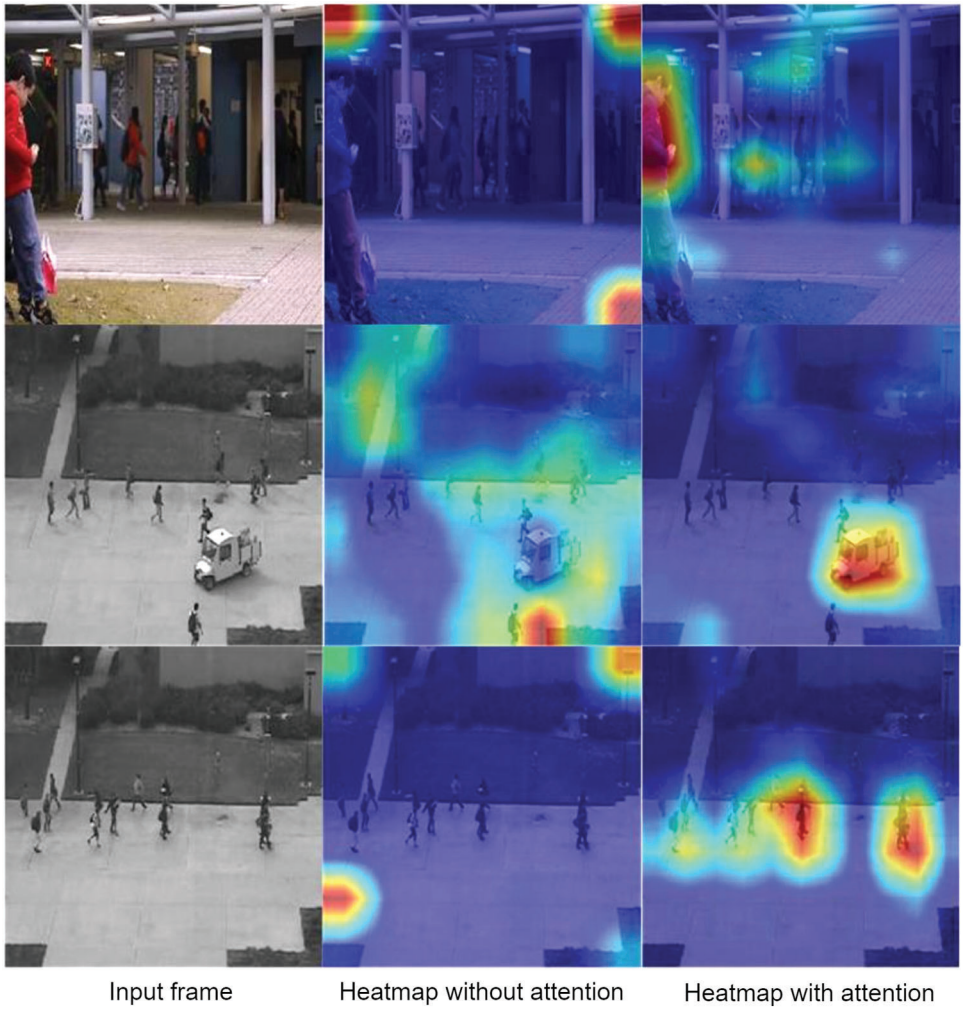

Figure 8: Feature heatmap of CCAE and BiAE on different datasets

Figs. 9 and 10 show ROC curves for UCSD Ped2 and CUHK Avenue datasets. As expected, CCAE outperforms BiAE for both datasets. The experimental results confirm that Criss-Cross attention module is effective, for its capability of capturing global dependencies during appearance encoding and temporal sequential learning.

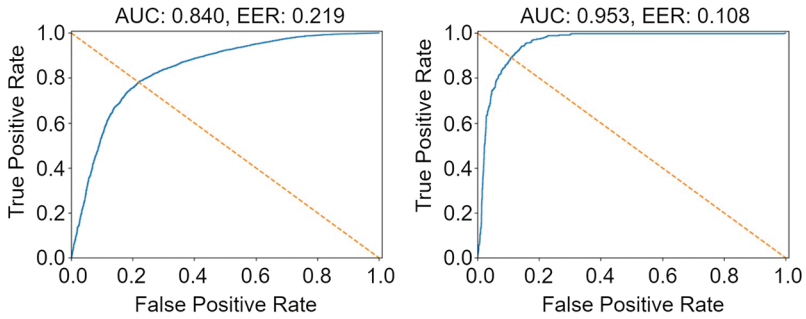

Figure 9: ROC curves of CCAE for Avenue and Ped2

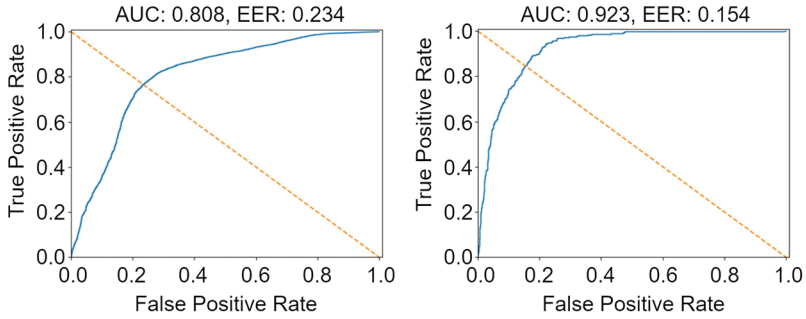

Figure 10: ROC curves of BiAE for Avenue and Ped2

5 Conclusions

In this paper, we presented an efficient unsupervised method CCAE for anomaly events detection by combining Criss-Cross attention and bi-directional ConvLSTM in autoencoder. The method employs Criss-Cross attention module in encoder for global appearance features, and a Criss-Cross attention module with bi-directional ConvLSTM for global temporal features. The two features are then fused by a decoder to reconstruct frames. In addition, an intensity loss and a gradient loss are designed to enhance normal pattern reconstruction. We perform extensive experiments on two public datasets UCSD Ped2 and CUHK Avenue and achieve competitive results with frame level AUC of 95.3% and 84.0% respectively. The experimental results validate the advantages of CCAE over other video anomaly detection methods. In the future, we consider investigating spatial and temporal features by graph convolutional neural networks to enhance feature representations of normal frames.

Funding Statement: This work was supported by the National Science Foundation of China under Grant No. 41971343.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Z. Xu, X. Zeng, G. Ji and B. Sheng, "Improved anomaly detection in surveillance videos with multiple probabilistic models inference," *Intelligent Automation & Soft Computing*, vol. 31, no. 3, pp. 1703–1717, 2022.

[2] B. R. Kiran, D. M. Thomas and R. Parakkal, "An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos," *Journal of Imaging*, vol. 4, no. 2, pp. 1–25, 2018.

[3] X. Duan, S. Ying, W. Yuan, H. Cheng and X. Yin, "A generative adversarial networks for log anomaly detection," *Computer Systems Science and Engineering*, vol. 37, no. 1, pp. 135–148, 2021.

[4] C. L. Li, K. Sohn and J. Yoon, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Nashville, Tennessee, USA, pp. 9664–9674, 2021.

[5] M. I. Georgescu, A. Barbalau and R. T. Ionescu, "Anomaly detection in video via self-supervised and multi-task learning," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Nashville, Tennessee, USA, pp. 12742–12752, 2021.

[6] W. Ullah, A. Ullah and I. U. Haq, "CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 16979–16995, 2021.

[7] L. Wang, F. Zhou and Z. Li, "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *25th IEEE Int. Conf. on Image Processing*, Athens, Greece, pp. 2276–2280, 2018.

[8] V. Mnih, H. Nicolas and G. Alex, "Recurrent models of visual attention," in *Proc. of the 27th Int. Conf. on Neural Information Processing Systems*, Montreal, Canada, pp. 2204–2212, 2014.

[9] A. Vaswani, "Attention is all you need," in *Proc. of the 31st Int. Conf. on Neural Information Processing Systems*, Long Beach, California, USA, pp. 6000–6010, 2017.

[10] Z. Huang, X. Wang and L. Huang, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Seoul, South Korea, pp. 603–612, 2019.

[11] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time MRF for detecting abnormal activities with incremental updates," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, Florida, USA, pp. 2921–2928, 2009.

[12] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[13] V. Mahadevan, W. Li and V. Bhalodia, "Anomaly detection in crowded scenes," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, California, USA, pp. 1975–1981, 2010.

[14] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 909–926, 2008.

[15] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Miami, Florida, USA, pp. 935–942, 2009.

[16] A. Adam, E. Rivlin and I. Shimshoni, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[17] Y. Fan, G. Wen and D. Li, "Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder," *Computer Vision and Image Understanding*, vol. 195, no. 102920, pp. 1–12, 2020.

[18] M. Sabokrou, M. Fayyaz and M. Fathy, "Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes," *Computer Vision and Image Understanding*, vol. 172, no. 5, pp. 88–97, 2018.

[19] F. Bi, X. Ma, W. Chen, W. Fang, H. Chen *et al.,* "Review on video object tracking based on deep learning," *Journal of New Media*, vol. 1, no. 2, pp. 63–74, 2019.

[20] Y. C. Su, "DAEN: Deep autoencoder networks for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4309–4321, 2019.

[21] M. Hasan, "Learning temporal regularity in video sequences," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, USA, pp. 733–742, 2016.

[22] O. Ronneberger, F. Philipp and B. Thomas, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, Cham, Springer, pp. 234–241, 2015.

[23] M. Sabokrou, M. Fathy and M. Hoseini, "Video anomaly detection and localization based on the sparsity and reconstruction error of auto-encoder," *Electronics Letters*, vol. 52, no. 13, pp. 1122–1124, 2016.

[24] K. Deepak, G. Srivathsan, S. Roshan and S. Chandrakala, "Deep multi-view representation learning for video anomaly detection using spatiotemporal autoencoders," *Circuits Systems, and Signal Processing*, vol. 40, no. 3, pp. 1333–1349, 2021.

[25] Y. S. Chong and H. T. Yong, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Int. Symp. on Neural Networks*, Hokkaido, Japan, pp. 189–196, 2017.

[26] W. Luo, W. Liu and D. Lian, "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1070–1084, 2021.

[27] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7132–7141, 2018.

[28] D. Gong, L. Liu and V. Le, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Seoul, South Korea, pp. 1705–1714, 2019.

[29] W. Sun, X. Chen, X. R. Zhang, G. Z. Dai, P. S. Chang *et al.,* "A multi-feature learning model with enhanced local attention for vehicle re-identification," *Computers, Materials & Continua*, vol. 69, no. 3, pp. 3549–3561, 2021.

[30] X. L. Wang, "Non-local neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, pp. 7794–7803, 2018.

[31] A. R. Sutanto and D. K. Kang, "A novel diminish smooth L1 loss model with generative adversarial network," in *Int. Conf. on Intelligent Human Computer Interaction*, Daegu, South Korea, pp. 361–368, 2020.

[32] S. Ren, K. He and R. Girshick, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.