Electronic Letters on Computer Vision and Image Analysis 24(1):134-152, 2025

Deep Learning-Based Video Anomaly Detection Using Optimised Attention-Enhanced Autoencoders

Anjali S* and Don S+

* *Dept. of Computer Science and Engineering,Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India*
+ *Dept. of Computer Science and Applications, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India*

Received 1st of December, 2024; accepted 1st of May 2025

Abstract

Anomaly detection in video is essential for applications like surveillance, healthcare, and industrial monitoring. Through the reconstruction of normal patterns and the computation of reconstruction error in relation to ground truth, convolutional autoencoders detect anomalies. Frames with errors above a threshold are flagged as abnormal. Existing approaches rely on fixed thresholds, which may not adapt well to varying lighting conditions, leading to false positives or missed anomalies. A novel autoencoder (SESAA) is proposed in this work that combines self-attention with squeeze-and-excitation (SE) blocks and improves video anomaly detection by using a thresholding technique for optimal threshold identification. Our adaptive thresholding technique leverages reconstruction cost, peak signal-to-noise ratio (PSNR) and frame brightness for optimal threshold identification, enhancing adaptability to different scenarios. Comparing with dynamic threshold methods, we assess our model using ROC and AUC metrics. Experiments on three benchmark datasets validate the efficacy of our method in precise anomaly detection through optimal thresholding.

Key Words: Computer Vision; Video surveillance; Optimal threshold detection; Autoencoder; Deep learning

1 Introduction

In computer vision, video anomaly detection is a method that automatically identifies odd or unexpected events in video footage. This has significant applications in a number of contexts, including traffic monitoring, smart homes, health care, and real-time video surveillance. Improving the monitoring system's dependability and security requires prompt anomaly detection. Because there are an infinite number of normal examples and a limited number of abnormal ones, anomaly identification is a difficult problem. Most of the existing methods identify anomalous occurrences as those that deviate from the patterns that are learnt from normal training data events [1, 2].

The two main categories of video anomaly detection techniques are handcrafted feature-based techniques [3, 4, 5, 6] and deep learning-based techniques [7, 8, 9, 10, 11, 12, 13, 14]. The first category includes several

Correspondence to: dons@am.amrita.edu

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2043
ELCVIA ISSN: 15108-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

manually created features that reflect normal patterns, such as object-wise trajectories, the histogram of magnitude and momentum (HOMM) [4], the histogram of optical flow (HOF) [6], and the histogram of orientated gradients (HOG) [5]. Applying these strategies to diverse settings is challenging since they require prior understanding of the scene. Deep learning techniques have been effectively applied to detect anomalous activity in videos because of the power of deep neural networks. Significant progress has been made in anomaly detection using prediction-based methods [15, 16] reconstruction-based deep learning techniques [2, 13, 11, 12] and their combination [17]. The basis for reconstruction techniques [8, 9, 10] is the idea that while anomalous events will have high reconstruction errors while normal events can be reliably reproduced. The prediction-based approaches [15, 16, 18] on the other hand, are based on the notion that abnormal occurrences cannot be accurately predicted, whereas normal events can be predicted with an acceptable level of efficiency.

In most of the existing works, the reconstruction error level is predefined in order to identify anomalies in video frames. The frames are deemed abnormal if the reconstruction errors are more than this threshold value; otherwise, they are regarded as normal [8]. However, these fixed threshold detection techniques will produce false positives because they cannot be adjusted for all conditions. Jia et al. [19] proposed a thresholding algorithm that adjusts the threshold to adapt illumination changes and evaluated their method using synthetic datasets. But this dynamic thresholding is highly sensitive to lighting conditions.

Attention in machine learning refers to a method that mimics the cognitive attention of humans by emphasizing certain portions of the input, such as an object, while ignoring other parts. When attention was focused on the foreground, where dynamic items were moving, rather than the static background area, performance in the anomaly detection domain was improved. For the purpose of identifying different video anomalous activities, a number of methods have been developed, including spatiotemporal auto-encoders [20, 21], deep convolutional auto-encoders [22], 3D convolutional auto-encoders [12] and convolutional long short-term memory (LSTM) auto-encoders [7]. The spatiotemporal convolutional networks often struggle with capturing both spatial and temporal features effectively, leading to blurred or inaccurate predictions. So convolutional autoencoders are combined with attention modules to increase reconstruction performance. ADANET, an autoencoder that combines a GAN model and an attention model, was proposed by Hao Song et al. [12]. Weichao Zhang et al. [23] proposed an auto-encoder that integrates dense residual networks and self-attention. For better video prediction, they combined gradient differences between frames and optical flow.

We propose an autoencoder, SESAA (Self-attention Enhanced Squeeze-and-Excitation Attention) to improve both spatial and temporal feature extraction by integrating self-attention mechanisms and squeeze-and-excitation (SE) blocks [24] into convolutional LSTM (ConvLSTM) networks. Additionally, a threshold optimisation algorithm is implemented that combines frame brightness and reconstruction cost from the autoencoder to provide a combined anomaly score. This algorithm finds the ideal threshold for anomaly categorisation using Youden's J statistic [25]. Results from experiments on benchmark datasets confirm the method's superiority over dynamic thresholding techniques, emphasising how it can be improved for more effective anomaly detection systems in real-time surveillance.

The main contributions of the paper are as follows:

- A novel autoencoder (SESAA) is proposed that combines self-attention with squeeze-and-excitation (SE) blocks into Convolutional LSTM (ConvLSTM) networks.
- An optimization algorithm that integrates different factors—reconstruction cost, peak signal-to-noise ratio (PSNR) and frame brightness—and forms a combined anomaly score.
- Comparison with dynamic thresholding methods that adjust the threshold based on statistical properties of the reconstruction cost and frame brightness.
- Evaluation on three benchmark video anomaly datasets using ROC and AUC metrics shows that the suggested optimization framework is effective at detecting anomalous events.

The rest of the document is organised as follows: The present state of anomaly detection research is reviewed in Section 2, which also offers details on deep learning frameworks and threshold detection methods. In Section

3, we present our proposed approach. The experimental results are discussed in Section 4. Section 5 covers the constraints of the proposed framework, and Section 6 concludes the article.

2 Related works

In computer vision, anomaly detection is one of the most challenging and persistent issues. Numerous public areas, including airports, shopping malls, subway stations, and train stations, have a large number of cameras installed due to the growing demand for public safety and surveillance. These cameras generate vast amounts of video data, which makes it challenging and time-consuming for a human operator to spot odd or suspicious activities. Consequently, significant efforts have been made in the direction of smart video surveillance, and numerous strategies have been put forth to enable notable advancements in the identification of anomalies in videos.

2.1 Deep learning models for video anomaly detection

The use of deep learning techniques to detect anomalous behaviour has advanced significantly. For anomaly detection, unsupervised [20, 26] or weekly supervised approaches [27, 28, 29] are used instead of supervised learning, which is unsuccessful because only trained events would be accurately detected. The two main categories of deep learning-based algorithms for visual anomaly identification are reconstruction-based and prediction-based approaches.

2.1.1 Reconstruction-based methods

One popular method for detecting anomalies in videos is the reconstruction-based method [7, 8, 9]. This method's concept is to utilize a training set of videos to create a model of normal behavior, which is then used to detect abnormalities in test videos. Reconstruction errors are assumed to be higher in frames with anomalies than in frames without anomalies. The autoencoder, which consists of a convolutional encoder, is the most widely used model.

2.1.2 Prediction-based methods

Frame prediction produces a high prediction error for samples with anomalies and attempts to anticipate regular occurrences accurately [15, 16, 18]. W. Liu et al. [15] used a video prediction system to address the anomaly detection issue. By contrasting the expected future time frame with its ground truth, they were able to identify an anomalous occurrence. Additionally, their method enforced a motion restriction to video prediction by introducing a continuous optical flow between the ground truth frames and the forecast frames. But prediction is not noise-resistant when applied to real-world datasets. Yao Tang et al. [17] created a system that strikes a compromise between the advantages of reconstruction- and prediction-based approaches. They developed a network that sequentially carries out frame prediction and reconstruction. Fig. 1 depicts the classification of approaches used for anomaly detection.

2.2 Threshold detection strategy

Despite the fact that reconstruction-based methods are good at identifying irregularities in videos, thresholding significantly affects how well they work overall. To categorize a frame as normal or abnormal, the majority of the existing works that are now in use have preset thresholds. Jia et al. [19] presented a dynamic thresholding method that dynamically adjusts the threshold to different light conditions using a pre-trained light module and automatically determines the threshold using only normal training sets. However, this dynamic thresholding approach may be sensitive to changes in frame brightness.

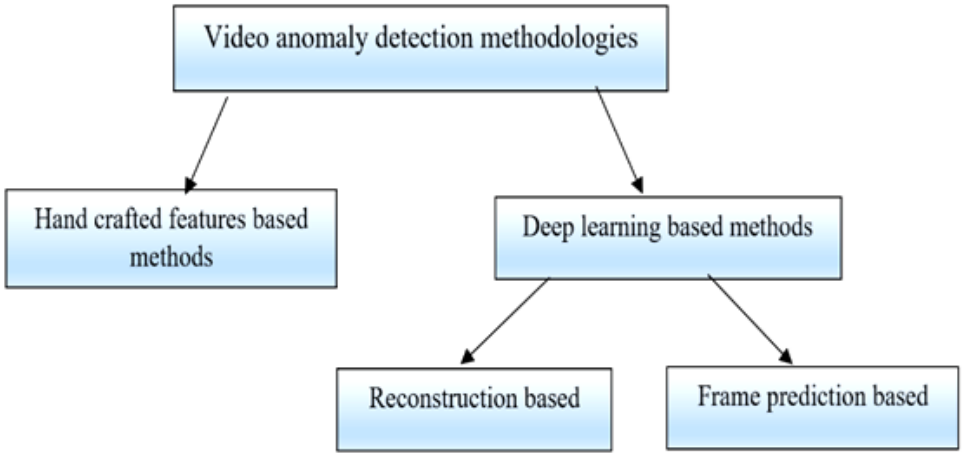


Figure 1: Anomaly detection techniques

3 Methodology

This section provides a detailed presentation of our video anomaly detection framework. As previously said, anomalous occurrences are extremely uncommon in everyday situations. As a result, gathering and labelling training data that includes all kinds of abnormalities is challenging. So we provide an unsupervised learning technique for identifying anomalous events in video in order to address this issue.

3.1 Data collections and preprocessing

The dataset used consists of surveillance videos collected from benchmark datasets like UCSDPed1 [30], UCSDPed2 [31] and the CUHK Avenue dataset [32]. The training video frames are divided into temporal sequences. A sliding window technique is implemented to generate sequences of sz (size of the window) consecutive frames. The video frames are normalized and resized to (256, 256, c), where c is 1 for grayscale images (UCSD Ped1, Ped2) and 3 for color images (Avenue).

3.2 Model development

We propose a network that uses the frame reconstruction method to detect video anomalies. An autoencoder framework that uses ConvLSTM backbone, squeeze-and-excitation (SE) block, and self-attention mechanism is developed for reconstruction. To identify patterns of regular activity from video recordings, it uses convolution, deconvolution, and Conv.LSTM layers. A popular network for spatiotemporal data reconstruction is the ConvLSTM network, which was introduced by Shi et al. [33]. ConvLSTM layers, which will identify temporal dependencies in the video sequences, form the basis of the model's core. It cannot, however, fully capture long-term dependencies. Hu et al. [24] presented SE blocks, which enhance channel-wise feature learning through adaptive recalibrating of channel-wise feature responses. In order to adjust channel-wise feature responses, SE blocks will be used after the first convolutional layers. By learning attention weights, SE blocks concentrate on the significance of each feature channel, enabling the network to suppress less valuable features and highlight more significant ones.

Applying self-attention to videos involves expanding the mechanism to handle temporal and spatial dimensions, which enables the model to recognize dependencies between different frames as well as inside individual frames. The Transformer architecture proposed by Vaswani et al. [34] made attention mechanisms popular, and they work well for modelling long-range dependencies in sequences, which makes them appropriate for

use with video data. In order to capture long-range dependencies and improve the model's capacity to focus on pertinent features, a self-attention layer will be added before the ConvLSTM layers.

Fig. 2 shows the suggested model's architecture. A video segment with 10 consecutive frames at a resolution of 256 * 256 serves as the network's input. The architecture consists of 9 layers, including two convolution layers, a ConvLSTM layer, two attention layers, two deconvolutional layers, one ConvLSTM (DeConv) layer, and one 1 x 1 convolution in the decoder part. Temporal and spatial encoders are included in the encoder section. Spatial and temporal decoders are included in the decoder also.

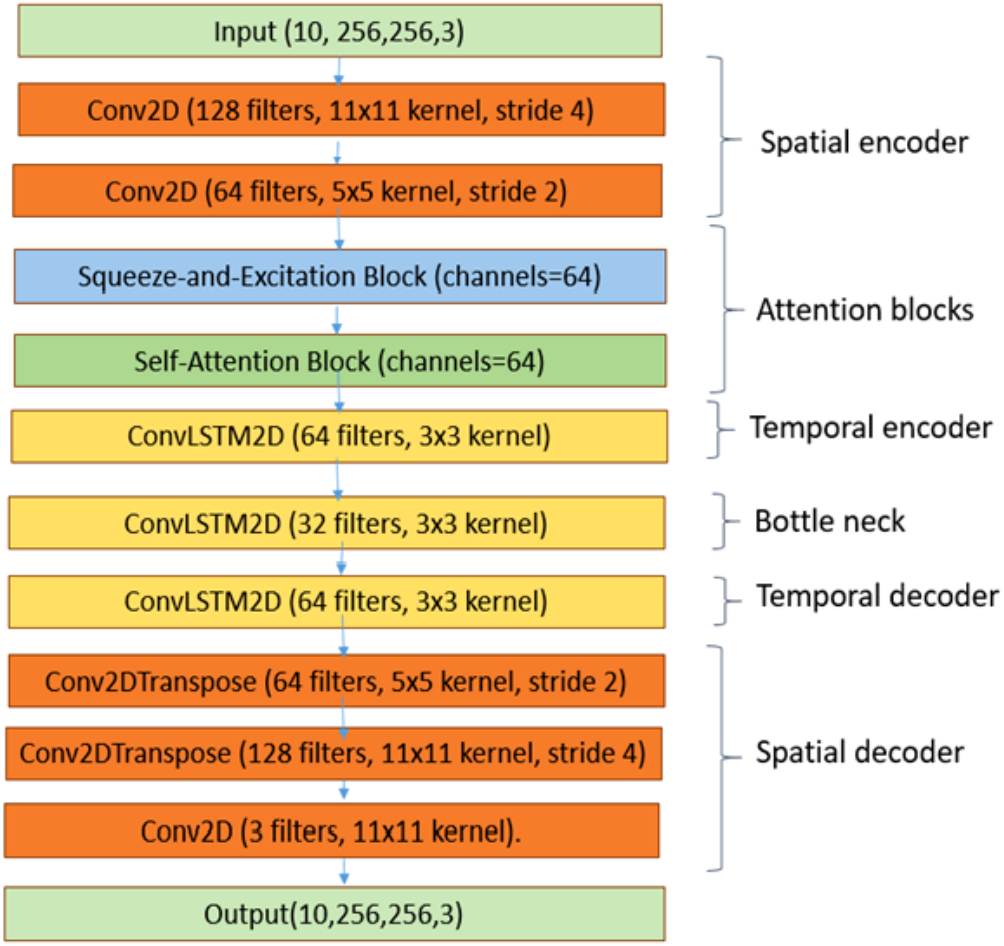


Figure 2: Proposed architecture

3.3 Model Training

The model is trained on normal video sequences to identify usual patterns of behaviour. The suggested model's training phase is depicted in Fig. 3. The input video segments are converted into spatial information using the convolution layers. The convolution process uses a sliding window to perform matrix multiplication between the picture patches and the filter. A reconstruction of the input sequence is the model's output when it gets an L-length video series as input. Each layer's output size is indicated by the numbers at the extreme right of Fig. 2. One frame at a time is fed into the spatial encoder, which processes L = 10 frames. The temporal encoder is then fed the concatenated encoded features of the ten frames in order to encode motion. The encoders and decoders mirror each other to rebuild the video volume.

The model's encoder and decoder components both make use of the ConvLSTM layer. ConvLSTM was

developed to record video sequences' temporal dynamics as well as spatial information. In addition to the ConvLSTM layer, the decoder part also contains deconvolutional or convolutional transpose layers for reconstruction. The hyper-parameters, including the number of kernels, kernel size, and strides, were empirically established prior to the kernel values being initialised at random. Using Mean Squared Error (MSE) as the loss function, the model is modified to lessen the discrepancy between the actual and predicted frames.

The testing phase utilizing the optimal thresholding approach is depicted in Fig. 4. The network receives a video segment as input that is made up of a series of grayscale or RGB images. Because the model is trained for normality, an abnormal frame will have higher MSE values, while a regular frame will have lower MSE values. A frame's normality score falls between 0 and 1. An optimal threshold (T) is determined using Youden's J statistic [25] utilizing the combined anomaly score, which is determined by a number of criteria, including reconstruction cost, frame brightness, and Peak Signal-to-Noise Ratio (PSNR). If the combined normality score (R) is less than threshold T, then it shows some abnormality in the frame.

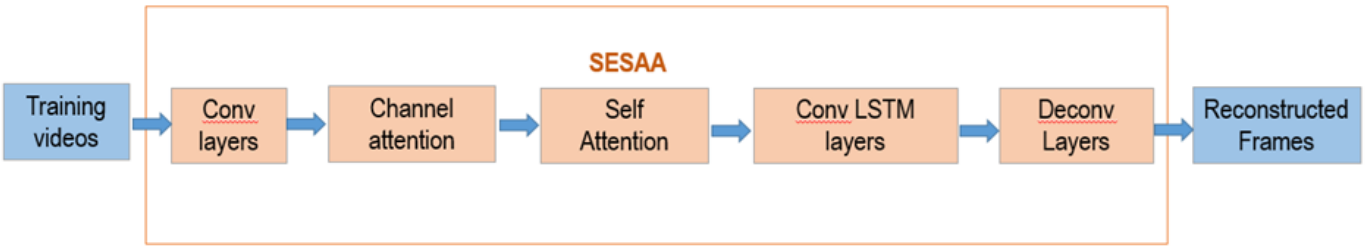


Figure 3: Training the autoencoder with self-attention and channel attention layers for video anomaly detection

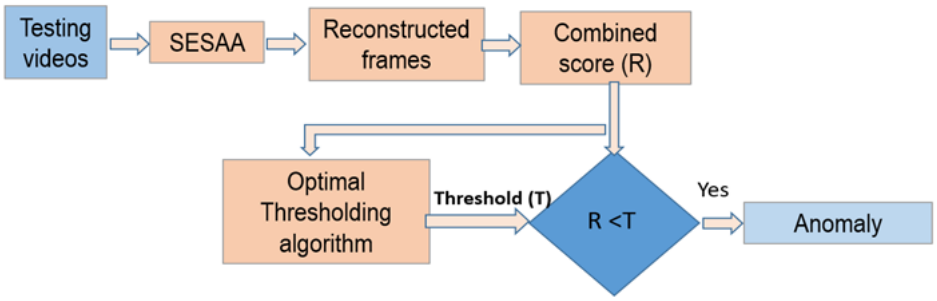


Figure 4: Testing phase of self-attention and channel attention autoencoder for video anomaly detection

3.4 Feature Extraction

During the training phase, no pre-trained network was utilized in order to extract features. The **reconstruction cost** E is calculated as the L2 norm between the original and reconstructed sequences, which is extracted for each video sequence using the trained autoencoder:

$$E = \|I - I_R\|_2 \tag{1}$$

where:

- I is the pixel intensity at a location in the original image,
- I_R is the intensity learned by the Convolutional Autoencoder.

The **sequence reconstruction cost** $sc(k)$ for n frames can be computed as:

$$sc(k) = \sum_{k=0}^{k+n} E(k) \tag{2}$$

Where $E(k)$ is the reconstruction cost for each frame at index k.

The **regularity score** s_r is computed by scaling between 0 and 1, and the **abnormality score** s_a is obtained by subtracting it from 1.

$$s_r(k) = \frac{sc(k) - sc(k)_{\text{max}}}{sc(k)_{\text{min}}} \tag{3}$$

The abnormality score is:

$$s_a = 1 - s_r \tag{4}$$

Peak Signal-to-Noise Ratio (PSNR) is a statistic used in anomaly detection to evaluate the quality of a reconstructed image. PSNR calculates the ratio of noise (reconstruction error) to the maximum signal (image intensity). Better image quality and a smaller reconstruction error are indicated by a greater PSNR. A Because the autoencoder model is unable to effectively rebuild the frame, a low PSNR indicates a high reconstruction error, which could indicate an anomaly in the video sequence. Because low PSNR values indicate possible irregularities, PSNR can be used to set a threshold for spotting anomalous sequences.

With the following formula, the **Peak Signal-to-Noise Ratio (PSNR)** is determined:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{\text{MSE}} \right) \tag{5}$$

Where:

- MAX_I represents the **maximum pixel intensity possible** in the image. The value of MAX_I is usually 255 for images that have pixel values between 0 and 255, such as 8-bit images. For normalized floating-point images, MAX_I is usually 1.0.
- The average squared difference between the original picture I and the reconstructed image $\hat{I}$ is quantified by **'MSE (mean squared error)**. It is described as:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (I_i - \hat{I}_i)^2 \tag{6}$$

Where:

- The pixel intensities at the i-th location in the original and reconstructed pictures are denoted as I_i and $\hat{I}_i$, respectively.
- N denotes the total number of pixels.

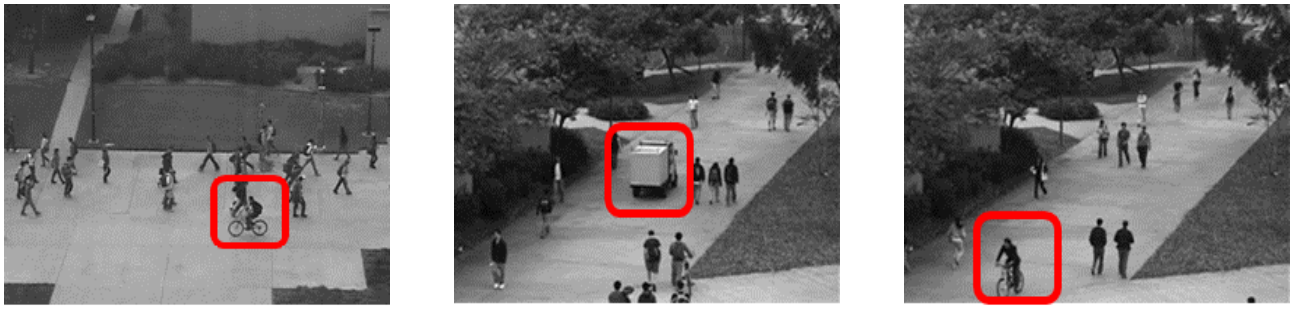
3.5 Threshold detection

To determine the optimal threshold, Youden's J statistic [25] is utilized, which maximizes the difference between TPR (true positivity rate) and FPR (false positivity rate). If the combined regularity score is less than the optimal threshold, frames will be identified as anomalous. The optimum threshold is compared to the dynamic threshold, which varies with each frame. The dynamic threshold is calculated using the normalised frame brightness as well as the standard deviation and mean of the combined score.

3.6 Evaluation

The effectiveness of our proposed method is evaluated using the area under the curve (AUC) and receiver operating characteristic (ROC) measures.

(a) Normal frames from the UCSD dataset.



(b) Abnormal frames from the UCSD dataset.

(c) Normal frames from the Avenue dataset.

(d) Abnormal frames from the Avenue dataset.

Figure 5: Normal and anomalous frames from benchmark datasets.

4 Experimental Results

4.1 Dataset description

Three benchmark video anomaly detection datasets are used in this paper: CUHK Avenue [32], UCSD Ped1 [30] and UCSD Ped2 [31]. Each dataset has a test set that contains both anomalous and normal frames and a training set that only contains normal films.

- **CUHK Avenue:**

  This dataset was made up of 21 test videos and 16 training videos, of 15,324 and 15,328 frames, respectively. 360 x 640 pixels was the resolution of every video frame. 47 unusual behaviors were captured, including throwing objects, going in the wrong way, and rushing across gates.

- **UCSD Ped1 dataset:**

  This dataset consists of 36 testing and 34 training video clips, each containing 200 frames. In the video, individuals are seen moving in groups to and from the camera. In this dataset, abnormal occurrences were characterized as people walking on the grass and the appearance of a wheelchair, skater, or cyclist.

- **UCSD Ped2 dataset:**

  The UCSD Ped2 dataset consists of 12 testing and 16 training video clips with varying frame counts. Pedestrians can be seen strolling parallel to the camera plane in the films. Wheelchairs, carts, bicycles, vehicles, and skaters are a few of the abnormalities discovered in the walkway. Fig. 5 shows some example frames of normal and abnormal examples of these three datasets.

4.2 Preprocessing and training

Three benchmark datasets—UCSD Ped1, UCPed2, and Avenue—are used to train the model. Videos are resized to 256 by 256 pixel resolution, and frames are normalized to fall between 0 and 1. To verify the adaptability of the model, both RGB and greyscale videos are used. Reconstruction error of the input volume is minimized to train the model. Here, we employ Adam Optimiser since it can automatically calculate the learning rate by using the model's weight update history. Regular video sequences are used for training in order to teach the model the patterns of usual behaviours.

4.3 Implementation details

The testing videos contains some anomalies which will not be reconstructed properly by the autoencoder and hence the regularity score will be low for such frames. The regularity score is calculated as the mean squared error between the original and reconstructed frames. Four different methods were explored in order to determine the optimum threshold, and it was found that the combined regularity score calculation and optimization produced the best results.

- **Method 1 : Statistical Thresholding for Anomaly Detection (STAD)**

  The regularity score's mean and standard deviation are used to determine the optimal threshold. The dynamic threshold stands in contrast to this. Experimental results on a variety of test videos demonstrate that anomalous frames cannot be correctly detected using this optimal threshold detection method. For UCSD Ped2 dataset, the normal frames where there is no cycle on the walkway are also detected as anomalies. Figure 6 displays the experimental findings on the UCSD Ped2 and Avenue datasets.

- **Method 2 : Gaussian-Smoothed Statistical Thresholding (GSST)**

A Gaussian smoothening filter is applied to the regularity scores and then optimal threshold is calculated using the standard deviation and mean of the regularity scores. As demonstrated on Fig. 6, the results are not better when compared to Method 1 for identifying the optimal threshold.

- **Method 3: Weighted PSNR Fusion with Youden's J Thresholding (W_ PSNR+YT)**

  Here, the PSNR value for every sequence is computed. (Anomalies are indicated by lower PSNR values). After that, by applying predetermined weights that normalize the brightness, PSNR, and reconstruction cost values, the combined score is calculated. The optimal threshold at which the AUC is maximized is then found using Youden's J statistic. As shown in Fig. 7, the results are better using the method of combined anomaly score generation and optimization using Youden's J statistic. These results are also contrasted with the dynamic threshold.

  For Test 32 of UCSD Ped1, as seen in Fig. 7, the low regularity score can be attributed to the bicycle on the walkway at the beginning. The score for regularity begins to rise after the bicycle departs. The regularity score drops immediately at the entry of a second bicycle at frame 60 and then rises again immediately upon its departure. By employing method 3, the best threshold was found to be 0.552, which allows for the accurate detection of abnormal bicycle entrance frames in pedestrian ways. As seen in Fig. 7, a person is throwing a bag, which is considered anomalous with respect to Test 5 of the Avenue dataset. After frame 30, the regularity score decreases, and as the person leaves, it increases. The optimal threshold obtained is 0.4.

- **Method 4 : Weighted Fusion with Youden's J Thresholding (WF+YT)**

  Here combined score is determined using normalized reconstruction cost and brightness with predefined weights. Here also Youden's J statistic is used to determine the optimal threshold for identifying anomalous frames. Similar to approach 3, the results are satisfactory here as well. Since PSNR values are somewhat sensitive to frame brightness, the PSNR calculation is not done in the combined score calculations here. For test 5 of the UCD Ped2 dataset, as shown in Fig. 8, a bicycle enters the pedestrian pathway at frame 60. As a result, the regularity score decreases, and it increases when the bicycle leaves. Fig. 8 displays the regularity scores obtained from the UCSD Ped1 dataset for Test Videos 32 and 31, UCSD Ped2 dataset for Test Videos 5 and 7, and the Avenue dataset for Test Videos 05 and 15. The AUC values for the bench mark datasets are depicted in Table 1.

	UCSD Ped1		**UCSD Ped2**		**Avenue**	
Method	**Optimal AUC**	**Adaptive AUC**	**Optimal AUC**	**Adaptive AUC**	**Optimal AUC**	**Adaptive AUC**
STAD	0.62	0.67	0.36	0.37	0.20	0.03
GSST	0.65	0.70	0.68	0.64	0.87	0.80
W_PSNR+ YT	0.77	0.77	0.65	0.68	0.80	0.76
WF+YT	**0.98**	**0.84**	**0.83**	**0.82**	**0.91**	**0.75**

Table 1: Comparison of area under the ROC curve (AUC) of different methods. A higher AUC is better.

4.4 Evaluation

The optimum threshold is compared with the dynamic threshold, and the irregularities are colored as shown in Fig. 8, which plots the regularity score curve. This study evaluated the model's performance using frame-level receiver operating characteristics (ROC) and its area under the curve (AUC) score, in line with previous methods [35], [7], and [20]. Separability is measured by the AUC score. It illustrates the model's capacity for class discrimination. A model that achieves an AUC score of 1 is perfectly able to differentiate the two classes. When the model has an AUC value of 0, it is reciprocating the findings, predicting a positive class as negative and vice versa. The model's inability to distinguish between classes is indicated by an AUC of 0.5. By

computing the TPR and FPR values at each classification threshold between 0 and 1, the ROC curve is plotted at every feasible threshold. The ROC curve for methods 3 and 4 of Section 4.3 is shown in Fig. 9. As compared to methods 1 and 2, the AUC values for methods 4 and 5 demonstrate a notable improvement, as seen in Table 1.

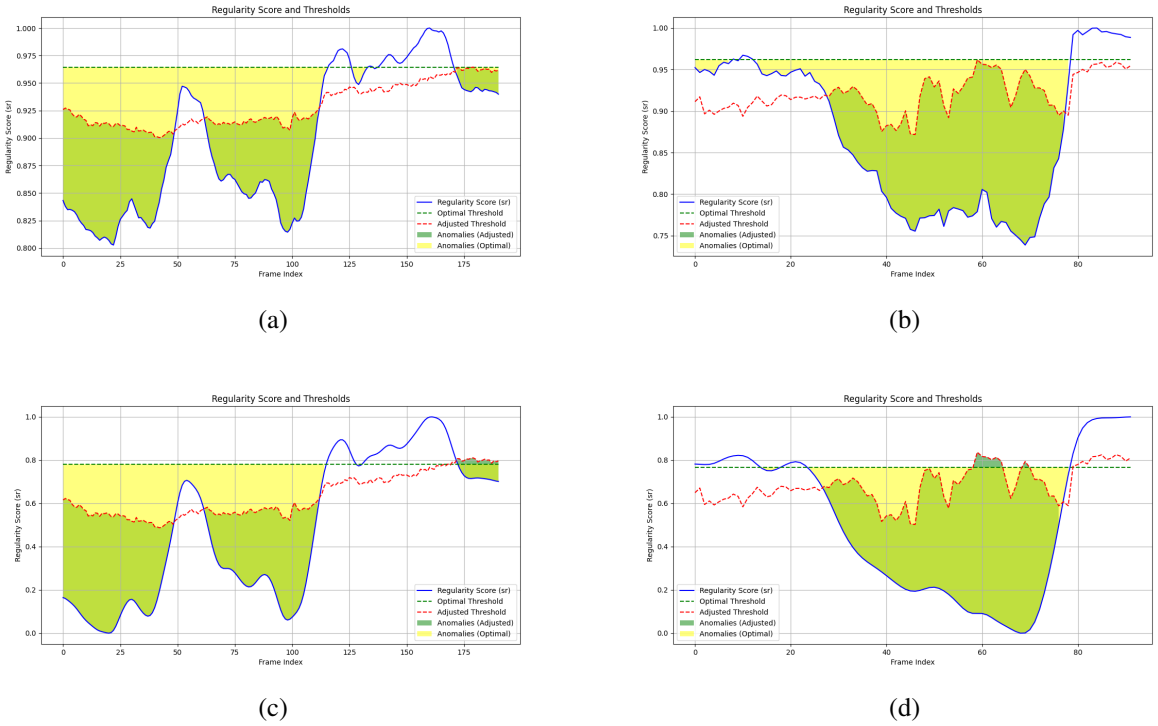


Figure 6: Regularity scores for STAD and GSST: (a) STAD for Test 32 of UCSD Ped1, (b) STAD for Test 05 of Avenue, (c) GSST for Test 32 of UCSD Ped1, (d) GSST for Test 05 of Avenue.

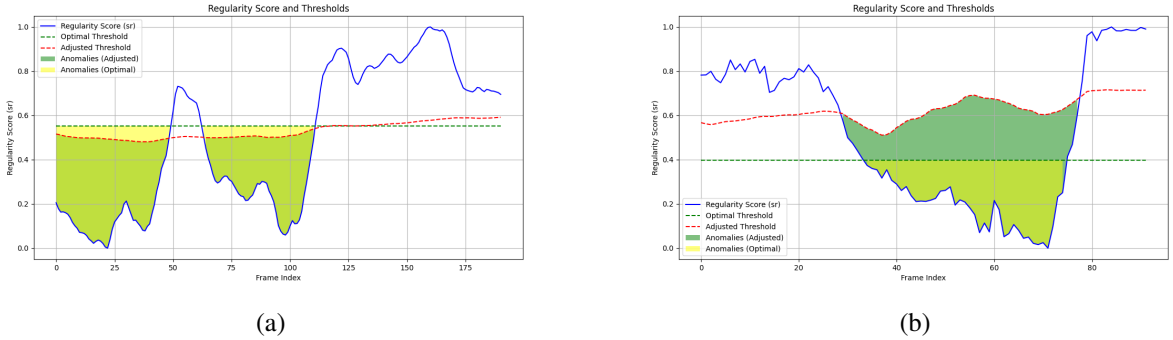


Figure 7: Regularity score for W_PSNR+YT: (a) Test 32 of the UCSD Ped1 dataset, (b) Test 05 of the Avenue dataset.

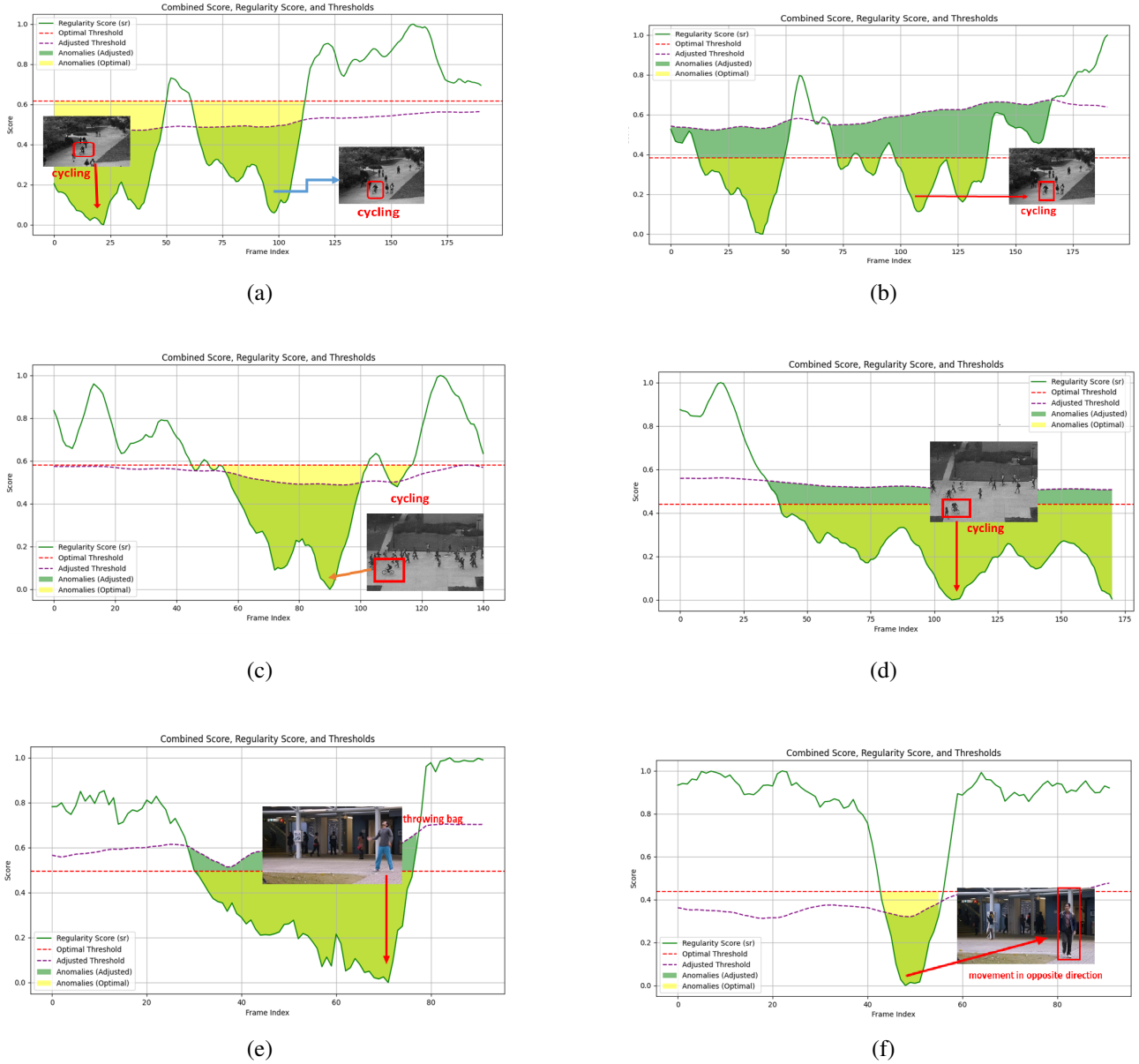


Figure 8: Regularity scores for Method 4 (WF+YT) (a) Test 32 of UCSD Ped1 (b) Test 31 of UCSD Ped1 (c) Test 05 of UCSD Ped2 (d) Test 07 of UCSD Ped 2 (e) Test 05 of Avenue (f) Test 15 of Avenue

4.5 Result Analysis

Our suggested method is effective in determining the optimum threshold to use for anomalous occurrence recognition, according to experiments conducted on three benchmark video anomaly detection datasets. The optimal thresholding is also contrasted with the AUC that is obtained by the dynamic thresholding approach. This suggests that the best way to identify anomalous frames is to compute the threshold using the combined anomaly score that was obtained from methods 3 and 4 of Section 4.3. It also demonstrates how illumination in real-time surveillance videos affects the dynamic threshold. Table 2 shows the optimal threshold obtained for three bench mark datasets using the four approaches mentioned in Section 4.3. The comparison of frame-level AUC% of popular eight reconstruction-based deep learning techniques is shown in Table 3. This demonstrates that the proposed method performs better in detecting anomalous occurrences than the most advanced

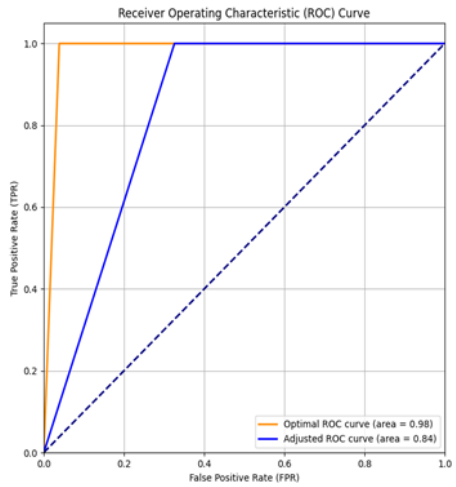
instruction-based deep learning systems.

Method	UCSD Ped1		UCSD Ped2		Avenue	
	Test 32	Test 31	Test 05	Test 7	Test 05	Test 15
STAD	0.96	0.91	0.99	0.82	0.97	0.90
GSST	0.78	0.64	0.81	0.64	0.79	0.66
W_PSNR+ YT	0.55	0.3061	0.42	0.51	0.40	0.59
WF+YT	0.61	0.38	0.50	0.44	0.50	0.43

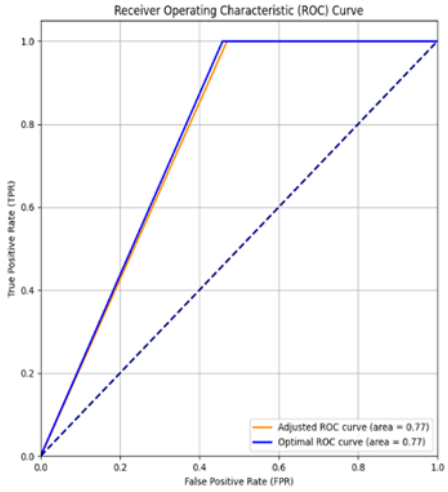
Table 2: Optimal thresholds obtained for the four approaches on benchmark datasets.

Method	UCSD Ped1	UCSD Ped2	Avenue
Chong et al. [20]	89.8	87.4	80.3
Hasan et al. [36]	81.0	90.0	70.2
Deepak et al. [7]	-	83.0	82.0
Nawaratne et al. [37]	75.2	**91.1**	76.8
Fang et al. [38]	95.6	86.3	73.2
Li et al. [39]	92.9	83.5	-
Song et al. [12]	90.3	90.4	89.2
Chang et al. [40]	96.5	86.0	73.0
Proposed (SESAA)	**98.0**	83.0	**91.0**

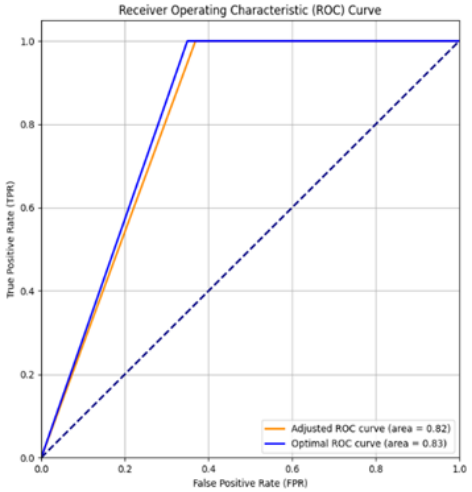
Table 3: Comparison of frame-level AUC of the different state-of-the art reconstruction-based deep learning methods in terms of AUC% on three benchmark datasets.Better performance is indicated by a higher AUC.



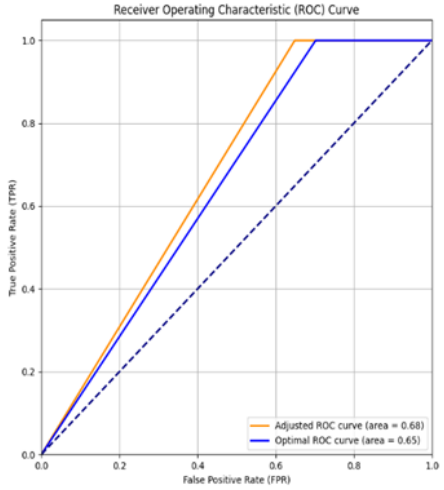
(a) UCSD Ped1: WF + YT



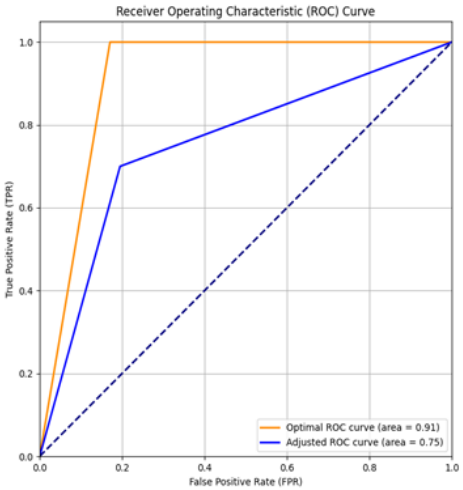
(b) UCSD Ped1: W_PSNR + YT



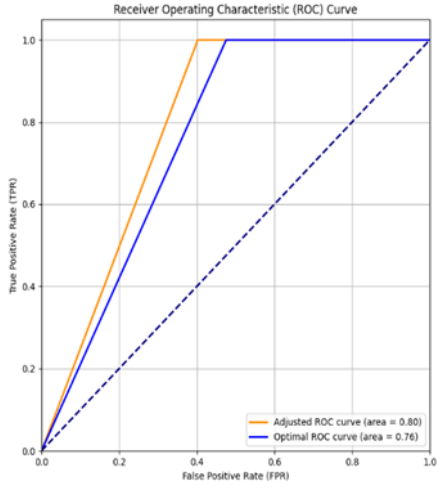
(c) UCSD Ped2: WF + YT



(d) UCSD Ped2: W_PSNR + YT



(e) Avenue: WF + YT



(f) Avenue: W_PSNR + YT

Figure 9: Optimal and adjusted ROC curves for the four methods across benchmark datasets: UCSD Ped1 (a, b), UCSD Ped2 (c, d), and Avenue (e, f).

5 Challenges

Some of the challenges faced in detecting anomalies in videos using the optimal thresholding technique are as follows:

- **Interpretability**: In some applications, it is crucial to comprehend the reason behind the selection of a specific threshold. Nevertheless, complicated computations or methods may be used during the threshold optimization process, making the outcome more difficult to understand.
- **Threshold Drift**: The ideal threshold may drift over time as a result of modifications to the underlying data distribution. Maintaining performance requires continuous calibration and monitoring, which can be resource intensive.
- **Computational Complexity**: Searching for an optimal threshold, especially in complex surveillance videos, can be computationally expensive. This is particularly challenging in real-time systems where quick decisions are required.

6 Conclusion

In this work, we proposed an autoencoder that enhances video anomaly detection by utilizing attention mechanisms and a thresholding technique for optimal threshold identification. The autoencoder combines self-attention with squeeze-and-excitation (SE) blocks. Reconstruction cost, frame brightness, and Peak Signal-to-Noise Ratio (PSNR) are the three criteria that are combined in this optimized framework to calculate thresholds automatically, allowing the system to adapt flexibly to various conditions. A comparison is made between this approach and dynamic thresholding techniques, which modify the threshold according to the statistical characteristics of the reconstruction cost and frame brightness. ROC and AUC measurements have been used to provide a thorough comparison and visualization of the findings. Our tailored method offers considerable advantages in selecting the optimum threshold for effective anomaly identification, as demonstrated by experimental results on three baseline datasets. With this technique, alerts for unusual activity can be generated in real-time monitoring by utilizing the optimal threshold.

References

[1] Le VT, Kim YG. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence. 2023 Feb;53(3):3240-54. https://doi.org/10.1007/s10489-022-03613-1

[2] Bajgoti A, Gupta R, Balaji P, Dwivedi R, Siwach M, Gupta D. SwinAnomaly: Real-Time Video Anomaly Detection Using Video Swin Transformer and SORT. IEEE Access. 2023 Oct 4. https://doi.org/10.36227/techrxiv.171174507.79352927/v1.

[3] ChenChen YT, Fang WH. Gaussian process regression-based video anomaly detection and localization with hierarchical feature representation. IEEE Transactions on Image Processing. 2015 Sep 17;24(12):5288-301.g KW, https://doi.org/10.1109/TIP.2015.2479561.

[4] Bansod SD, Nandedkar AV. Crowd anomaly detection and localization using histogram of magnitude and momentum. The Visual Computer. 2020 Mar; 36(3):609-20. https://doi.org/10.1007/s00371-019-01647-0.

[5] Kaltsa V, Briassouli A, Kompatsiaris I, Hadjileontiadis LJ, Strintzis MG. Swarm intelligence for detecting interesting events in crowded environments. IEEE transactions on image processing. 2015 Mar 6; 24(7):2153-66. https://doi.org/10.1109/TIP.2015.2409559.

[6] Wang T, Qiao M, Zhu A, Niu Y, Li C, Snoussi H. Abnormal event detection via covariance matrix for optical flow based feature. Multimedia Tools and Applications. 2018 Jul; 77:17375-95. https://doi.org/10.1007/s11042-017-5309-2.

[7] Deepak K, Chandrakala S, Mohan CK. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing. 2021 Feb;15(1):215-22. https://doi.org/10.1007/s11760-020-01740-1.

[8] Wang Y, Qin C, Bai Y, Xu Y, Ma X, Fu Y. Making reconstruction-based method great again for video anomaly detection. In2022 IEEE International Conference on Data Mining (ICDM) 2022 Nov 28 (pp. 1215-1220). IEEE. https://doi.org/10.1109/ICDM54844.2022.00157.

[9] Zhao Y, Deng B, Shen C, Liu Y, Lu H, Hua XS. Spatio-temporal autoencoder for video anomaly detection. In Proceedings of the 25th ACM International Conference on Multimedia 2017 Oct 23 (pp. 1933-1941). https://doi.org/10.1145/3123266.3123451.

[10] Liu Y, Liu J, Lin J, Zhao M, Song L. Appearance-motion united auto-encoder framework for video anomaly detection. IEEE Transactions on Circuits and Systems II: Express Briefs. 2022 Mar 22; 69(5):2498-502. https://doi.org/10.1109/TCSII.2022.3161049.

[11] Sun C, Jia Y, Song H, Wu Y. Adversarial 3d convolutional auto-encoder for abnormal event detection in videos. IEEE Transactions on Multimedia. 2020 Sep 10; 23:3292-305. https://doi.org/10.1109/TMM.2020.3023303.

[12] Song H, Sun C, Wu X, Chen M, Jia Y. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. IEEE Transactions on Multimedia. 2019 Nov6;22(8):2138-48. https://doi.org/10.1109/TMM.2019.2950530.

[13] Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, Hengel AV. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In-Proceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 1705-1714). https://doi.org/10.1109/ICCV.2019.00179.

[14] Doshi K, Yilmaz Y. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. Pattern Recognition. 2021 Jun 1;114:107865. https://doi.org/10.1016/j.patcog.2021.107865.

[15] Liu W, Luo W, Lian D, Gao S. Future frame prediction for anomaly detection—a new baseline. In Proceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 6536-6545). https://doi.org/10.1109/CVPR.2018.00684.

[16] Saypadith S, Onoye T. Video anomaly detection based on deep generative network. In2021 IEEE International Symposium on Circuits and Systems (ISCAS) 2021 May 22 (pp. 1-5). IEEE. https://doi.org/10.1109/ISCAS51556.2021.9401642.

[17] Tang Y, Zhao L, Zhang S, Gong C, Li G, Yang J. Integrating prediction and reconstruction for anomaly detection. Pattern Recognition Letters. 2020Jan1;129:123-30. https://doi.org/10.1016/j.patrec.2019.11.024.

[18] Li C, Li H, Zhang G. Future frame prediction based on generative assistant discriminative network for anomaly detection. Applied Intelligence. 2023 Jan;53(1):542-59. https://doi.org/10.1007/s10489-022-03488-2.

[19] Jia D, Zhang X, Zhou JT, Lai P, Wei Y. Dynamic thresholding for video anomaly detection. IET Image Processing. 2022 Sep;16(11):2973-82. https://doi.org/10.1049/ipr2.12532.

[20] Chong YS, Tay YH. Abnormal event detection in videos using spatiotemporal autoencoder. InAdvances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part II 14 2017 (pp. 189-196). Springer International Publishing.https://doi.org/10.1007/978-3-319-59081-3_23.

[21] Mangai P, Geetha MK, Kumaravelan G. Two-Stream Spatial–Temporal Feature Extraction and Classification Model for Anomaly Event Detection Using Hybrid Deep Learning Architectures. International Journal of Image and Graphics. 2023 Jul 8:2450052. https://doi.org/10.1142/S0219467824500529.

[22] Ribeiro M, Lazzaretti AE, Lopes HS. A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recognition Letters. 2018 Apr 1;105:13-22. https://doi.org/10.1016/j.patrec.2017.07.016.

[23] Zhang W, Wang G, Huang M, Wang H, Wen S. Generative adversarial networks for abnormal event detection in videos based on self-attention mechanisms. IEEE Access. 2021 Sep 7; 9:124847-60. https://doi.org/10.1109/ACCESS.2021.3110798.

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 7132-7141). https://doi.org/10.1109/CVPR.2018.00745.

[25] Perkins NJ, Schisterman EF. The Youden Index and the optimal cut-point corrected for measurement error. Biometrical Journal: Journal of Mathematical Methods in Biosciences. 2005 Aug;47(4):428-41. https://doi.org/10.1002/bimj.200410133.

[26] Tao Y, Hu Y, Chen Z. Memory-guided representation matching for unsupervised video anomaly detection. Journal of Visual Communication and Image Representation. 2024 May 1;101:104185. https://doi.org/10.1016/j.jvcir.2024.104185.

[27] Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp.6479-6488). https://doi.org/10.1109/CVPR.2018.00678.

[28] Lv H, Zhou C, Cui Z, Xu C, Li Y, Yang J. Localizing anomalies from weakly-labelled videos. IEEE transactions on image processing. 2021 Apr 19; 30:4505-15. https://doi.org/10.1109/TIP.2021.3072863.

[29] Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. InProceedings of the IEEE/CVF international conference on computer vision 2021 (pp. 4975-4986). https://doi.org/10.1109/ICCV48922.2021.00493.

[30] Mahadevan V, Li W, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. In2010 IEEE computer society conference on computer vision and pattern recognition 2010 Jun 13 (pp. 1975-1981). IEEE. https://doi.org/10.1109/CVPR.2010.5539872.

[31] Li W, Mahadevan V, Vasconcelos N. Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence. 2013 Jun 13;36(1):18-32. https://doi.org/10.1109/TPAMI.2013.111.

[32] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab. InProceedings of the IEEE international conference on computer vision 2013 (pp. 2720-2727). https://doi.org/10.1109/ICCV.2013.338.

[33] Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems. 2015;28.

[34] Ashish V. Attention is all you need. Advances in neural information processing systems. 2017;30:I. https://doi.org/10.1145/3347146.3359342.

[35] Jin P, Mou L, Xia GS, Zhu XX. Anomaly detection in aerial videos with transformers. IEEE Transactions on Geoscience and Remote Sensing. 2022 Aug 11;60:1-3. https://doi.org/10.1109/TGRS.2022.3198130.

[36] Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS. Learning temporal regularity in video sequences. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 733-742). https://doi.org/10.1109/CVPR.2016.86.

[37] Nawaratne R, Alahakoon D, De Silva D, Yu X. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. IEEE Transactions on Industrial Informatics. 2019 Aug 29;16(1):393-402. https://doi.org/10.1109/TII.2019.2938527.

[38] Fang Z, Zhou JT, Xiao Y, Li Y, Yang F. Multi-encoder towards effective anomaly detection in videos. IEEE Transactions on Multimedia. 2020 Nov 18;23:4106-16. https://doi.org/10.1109/TMM.2020.3037538.

[39] Li N, Chang F, Liu C. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Transactions on Multimedia. 2020 Apr 2;23:203-15. https://doi.org/10.1109/TMM.2020.2984093.

[40] Chang Y, Tu Z, Xie W, Luo B, Zhang S, Sui H, Yuan J. Video anomaly detection with spatio-temporal dissociation. Pattern Recognition. 2022 Feb 1;122:108213. https://doi.org/10.1016/j.patcog.2021.108213.