Journal of Theoretical and Applied Information Technology
31st December 2023. Vol.101. No 24

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BILAE-GAN FRAMEWORK FOR ANOMALY DETECTION IN VIDEO SURVEILLANCE

SWAPNA.C[1], DR.B.PADMAJA RANI [2]

[1]Asst. Prof., AI&DS Department, SCTEW, India

[2]Prof., CSE Department, JNTUH, India.

[1]swapnac.jntuh@gmail.com; [2]padmaja_JNTUH@jntuh.ac.in

ABSTRACT

In recent years, increasing the use of surveillance cameras with less manpower makes automatic video surveillance systems to become more important. Recent advances in video anomaly identification have mostly focused on improving performance with available datasets. We propose a Bidirectional Long-Short term memory-based Convolutional Autoencoder Generative Adversarial Network (BiLAE-GAN) method for video surveillance. During training the model learns the normal data distribution of data in the Generator and the detection of anomalies in the discriminator. Bidirectional Long-Short term network in Convolutional Autoencoder in Generator for reconstruction, Encoder features of Generated image and real image to discriminator to identify Anomaly. At the anomaly detection phase, anomalies are identified based on reconstruction error and discrimination results. Our proposed method validation benchmark datasets such as UCSD Ped1, UCSD Ped2, and CHUCK Avenue dataset with performance metrics AUC, EER.

Keywords: *BiLAE-GAN, Encoder-Decoder, BILSTM, Svdloss, Advloss*

1. INTRODUCTION

The widespread use of closed-circuit television (CCTV) cameras has resulted in an enormous amount of real-time video data. The manual evaluation by human beings is no longer feasible. Detecting anomalous behavior in video surveillance is critical to public safety. Until now, video surveillance systems have been progressively distributed throughout our entire society, and the volume of surveillance video data has grown rapidly.

Our Proposed System is useful in various areas like Human tracking results are further exploited to detect suspicious behaviors such as entering a secured place, running or moving around capriciously, loitering against traffic, dropping any suspicious things in public places. Other suspicious actions due to loitering such as drug-dealing, bank robbery, theft and picpocketing can be prevented by activity recognition and prediction approaches. Further, vision-based surveillance systems are more attractive and authenticative since it can be performed at a distance and secretly, whereas other biometric methods would require physical touch or close distance recognition. In entertainment environment, various events can be recognized. Safety can be assured in swimming pools.

Despite significant research in intelligent video surveillance, the potential for continual learning from new data remains untapped While current anomaly detection approaches perform well on benchmark datasets such as UCSD Pedestrian1, UCSD Pedestrian2 and CUHK Avenue, advancement in this field has stalled.

There are 3 types of approaches for video Anomaly Detection 1. Supervised, Semi Supervised, Unsupervised. Supervised Approach uses Labeled Data to train model. Semi-supervised anomaly detection approaches are broadly classified into two types: reconstruction-based techniques and prediction-based techniques. Previous reconstruction methods [1], [2], [4], [5], [9] relied on hand-crafted appearance and motion features. Then, using those features, a dictionary is acquired to sparsely encode all normal occurrences with minor reconstruction errors. During testing, the features correlating to anomalous occurrences may cause a high reconstruction mistake. However, the optimization of sparse coefficients can be extremely time-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

consuming, and the efficacy of anomaly detection is regulated by hand-crafted features, and These two factors limit the dictionary learning approach. Later, as deep learning advanced, various studies [6]-[8], [10]-[12],[23],[24] employed unsupervised approach to autonomously identify deep features rather than hand-crafted features. Hasan et al. [8] proposed using an auto-encoder to reconstruct typical occurrences with minor reconstruction errors. Because of its ability to represent high-dimensional image data, Generative Adversarial Networks (GANs) have recently become the state of the art in anomaly identification.

Generative Adversarial Networks consist of two different networks, a generator and a discriminator, both trained with unlabeled data.The generator *Ge* aims to capture the data distribution and generate realistic video frames, by building a data distribution for the input data *Xi* via a mapping from a prior latent space noise distribution *z*. The objective of the discriminator *Di*, instead, is to find the probability of the sample being outputted by the generator. Generator and discriminator compete against each other by playing a zero-sum min-max game: $\min_{Ge} \max_{Di} V(Di, Ge) = E_{X\ p}data(Xi)$ log $Di(Xi) + E_{Z\ p}z(Z)$ log(1 $Di(Ge(Z))$). Most unsupervised GAN architectures use shallow networks that are meant to learn only spatial characteristics while neglecting the critical temporal component of videos.

Contributions:

Our proposed model BILAE-GAN (Bidirectional LSTM Auto Encoder Generative Adversarial Networks) can identify anomalous easily than Less parameterized SVD-GAN architecture scored well on benchmark datasets, however model training requires more iterations. Less iterations are needed to train the model with our suggested architecture [4].

- The use of Conv BiLSTM layers in the Auto Encoder structure of the Generator with forward and backward directions maintains a prior history, which increases learning and hence the generator's reconstruction capabilities.
- A generator structure based on our own time distributed depth-wise convolution layers, resulting in greater efficiency without sacrificing feature extraction, providing a model that is both lightweight and more efficient.

2. RELATED WORK

The aim of an anomaly detection system is to predict and prevent anomalous (criminal) behaviour. In [20], Liu et al. introduced a future frame prediction framework (FFP) for anomaly identification, inspired by the good efficiency of the video prediction model in [19]. The frame discriminator's role is to compare the predicted future frame with the actual future frame to determine whether the input is from a real distribution or is generated by the generator, which can effectively enhance the generator's durability and the accuracy of the predicted frame via an interactive game with the generator.

To encourage the generator to predict video frames that are consistent with the real sequence in the temporal connection, a sequence discriminator is employed to support the temporal consistency of the predicted frame in [21]. This tool determines whether the frame sequence contains false frames or not. Motivated by the research in [22]. An end-to-end trainable bidirectional retrospective generative adversarial network is available for video anomaly detection. The prediction network may more completely exploit the bidirectional mapping relationship between video frame sequences by implementing the bidirectional prediction paired with the retrospective prediction in [25]. This will lead to more accurate future-frame predictions of typical events.

In an encoder-decoder model for frame prediction and anomaly detection by reconstruction, Jefferson et al. developed a Conv-LSTM network [26]. The same design was shown to be effective for detecting video anomalies [27]. The convolutional LSTM is then used to extract features from the input video frames before utilizing deconvolution to rebuild the video frames. In fact, LSTM autoencoders are well-suited for extracting spatial-temporal data. Shi et al. [28] and Patraucean et al. [29] used multilayer convolutional LSTMs in an autoencoder architecture for feature extraction in video sequence data. Conv-LSTMs exhibit a clear capacity to forecast the next frames and can extract both spatial representations and spatiotemporal data from a series of video frames.

Anomaly localization accuracy and false-positive anomaly detection are improved by.DR-cGAN. The somewhat lengthy computation time and

need for numerous training samples of typical occurrences are two drawbacks of DR-cGAN[1]. Our proposed system take less time to train the Model with available train data. Simple gaussian models cannot handle the complexity of real data, whereas ConvLSTM-VAE models have an advantage in terms of training time [2].ConvBiLSTM layer of our proposed model can handle real data and less time to train model.

Encoder-decoder LSTM [34] is presented for unsupervised learning. Spatiotemporal networks (STNs) are becoming popular for learning spatial and temporal features [35], where RNNs and CNNs extract spatiotemporal data concurrently for anomaly detection. The ConvLSTM [36] is another model in which a convolutional layer filters the output of CNNs before feeding it to an LSTM. A convolutional layer, an alternative to the fully connected layer in LSTM, drastically reduces the number of parameters. When GRU is used instead of LSTM in ConvLSTM, the parameters are reduced even further. The parameters have been reduced by 25%.

The capacity to train a deeper network with fewer parameters and less memory use is provided by multi-scale U-Net architecture. A scene's ambiguous anomalous objects cannot be distinguished by limitation [3].

3. PROPOSED VIDEO SURVEILLANCE ANOMALY DETECTING SYSTEM

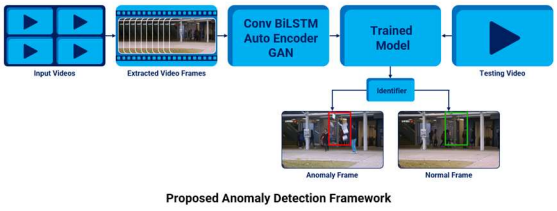

Figure 1: Proposed video Surveillance Anomaly Detection System

This paper presents a unique frame prediction framework as model for anomaly detection known as a Bidirectional Long-Short term memory-based Convolutional Autoencoder Generative Adversarial Network (**BiLAE-GAN**) that overcomes these restrictions. The **BiLAE-GAN** comprises a generator and a discriminator that are both made up of 2-Dimensional (2D) Bidirectional convolutional LSTM autoencoder networks that can be trained end-to-end.

The overall architecture of Bidirectional Long-Short term memory-based Convolutional Autoencoder Generative Adversarial Network (**BiLAE-GAN)** is shown in Figure 1. This framework has the following traits:

1) It can thoroughly investigate the bidirectional mapping relationship between video frame sequences in order to more precisely establish the mapping model from certain frames in the past to a reconstruction of a future frame for normal events;
2) The motion constraint can be carried out from the perspective of long-term temporal consistency in order to ensure that the predicted frame and the actual frame coincide with normal events, in terms of the motion.

3.2 Network Architecture

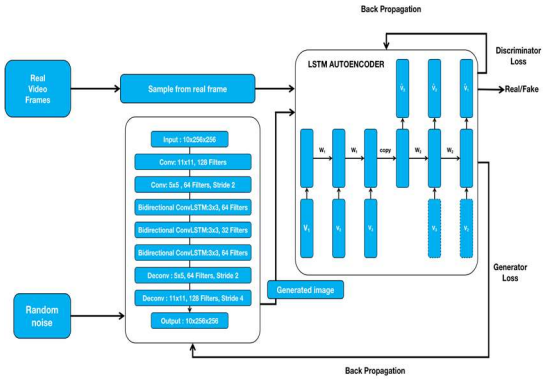

Figure 1 Proposed BiLAE-GAN Architecture

3.1 Video Pre-Processing

Reading all videos in folder and dividing video into frames at the rate of 24 frames for second. Resizing frame and Reading difference between each frame and its neighboring frame is computed in order to take spatial motion and loading frames stride by stride. In Conv-LSTMs the amount of information from the previous time step received by the hidden state is partly determined by the size of convolutional filter in the hidden-to-hidden connection. To capture faster motions large transitional kernels are used, while for slower motions small kernels can do [41]. GAN performance, however, degrades drastically as the number of parameters increases, leading GANs to often fail on high dimensional data. To overcome this, in our model *depth-wise separable convolutions* are used within the ConvBiLSTM, reducing the size of the model and the chance of overfitting during GAN training. Depth wise separable convolution is followed by pointwise convolution to deal with the spatial and depth dimensions of the video frames, thus splitting a 3×3 kernel into 3×1 and 1×3 kernels. Each input frame is processed using three separate filters for the R, G and B channels.

We specifically suggest a Bidirectional ConvLSTM autoencoder to properly exploit the bidirectional mapping link among video frame sequences. Because of the bidirectional ConvLSTM, the generator can forecast both the future and the past. In the bidirectional prediction process, backward prediction refers to predicting a previous frame by watching present frames, and forward prediction refers to predicting a future frame by watching present frames. Then, in order to execute prediction for review verification, the forward forecasted frame and the backward predicted frame are separately added to the matching input sequence. By losses and gradient loss between the predicted or reconstrcuted frame and the actual frame, the appearance constraint is then applied.

The proposed method tackles two key issues preventing a GAN-based architecture from reliably reconstructing video frames. Overfitting is addressed by reducing the number of parameters via depth-wise separable convolution.

3.2.1. Generator

Our proposed architecture (Figure 1) is based on the generative adversarial network principle and uses an encoder-decoder-encoder pipeline as the Generator with Time Distributed separable and Bidirectional ConvLSTM layers, BatchNormalization, tanh activation function which learns feature representations with LeakyReLU directly from the input samples, and an encoder-based.

At each time step, a batch of fixed-duration video frames is passed as input to both the generator G and the discriminator D. Each input image X is passed to the encoder E in the generator G, which maps it to a latent space $Z = E(X)$. The decoder then transposes these latent space vectors back to the image data space, implementing a mapping. The first contribution of this paper is an original Generator architecture, shown in Figure 2, where it can be seen how the generator learns the temporal dependencies within a video sequence using temporal blocks of frames.

At training time, the network learns the joint posterior distribution of the data $Ge(Z, Xi)$, and each input sample Xi is encoded using its latent representation. Training is performed using the inverse mapping from image data to latent space proposed by Lipton et al [18].

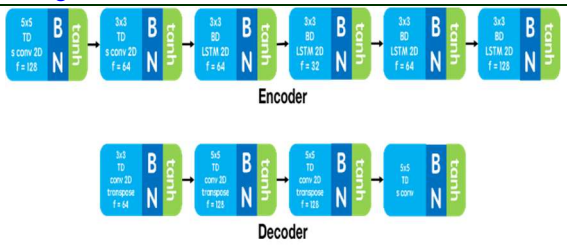

Figure 2 Encoder and Decoder Structure

3.2.2 discriminator

We also incorporate a frame discriminator to detect whether the image created by the generator is real or false, which deters the generator from producing blatantly fraudulent images and enhances the resilience of the generator and the quality of the predicted or reconstructed frames. Furthermore, we suggest a discriminator built of 2D convolutional neural networks to capture the long-term temporal information between video frame sequences in order to make the predicted or reconstructed frame consistent with the real object in motion.

Discriminator with Time Distributed Dense layer, GlobalAveragepooling2D, Sigmoid Function, which aims to discern real from fake images. The generated image $\bar{X} = Ge(X)$ and the input image X are then passed to the discriminator, which discriminates real from generated frames using a sigmoid activation function and backpropagates the loss to the generator itself.

3.2.3 losses

By calculating losses and updating weights of model, we trained our model efficiently.

SVD loss Our proposal is to minimise the empirical expectation of the $L2$ norm of the difference between the low-rank SVD approximations of the input image X and of the generated image $Ge(Xi)$, which we term

$$SVDloss = \mathrm{E}X\sim\mathrm{P}(X)\ |Xbr - Ge[(Xi)r\ | \quad (1)$$

where Xbr and $Ge[(X)r$ are the rank-r approximations of X and $Ge(Xi)$, respectively. Our conjecture, which we empirically validate in this paper, is that minimising the SVD loss is indeed correlated with minimising the KL divergence between real and generated data, and should thus have positive effects on the convergence of our model. The use of the original SVD_{loss} in the generator allows us to minimize the distance between low-rank optimal approximations of input and generated images, aiding network convergence. As a result, our Bidirectional Long-Short term memory-based Convolutional Autoencoder Generative Adversarial Network (BiLAE-GAN) model by using Bidirectional LSTM reconstructed image within less iterations comparing with LSTM.

The *adversarial loss ADVloss* is the Jensen-Shannon divergence of the output of the discriminator *Di* for the input image *X* and the corresponding generated image *Ge*(*Xi*) [32], namely:

$$ADVloss = E_{X\sim P(X)} |Di(Xi) - Di(Ge(Xi))|. \quad (2)$$

A *contextual loss* based on the *L*1 norm is used for penalising the distance between the input image *X* and the reconstructed image *Ge*(*Xi*) [13]:

$$CONTloss = E_{X\sim P(Xi)} |Xi - Ge(Xi)| \quad (3)$$

SVD loss can thus be seen as a an efficient form of contextual loss, based on the *L*2 norm. Compared to the *L*1 norm, *L*2 better penalises out liers, as deviations are magnified by taking the square.

Finally, the *encoder loss* minimises the distance between the bottleneck features of the input $Z = E(Xi)$ and the encoded features of the generated image.

In our model, more stable GAN training is achieved by using as overall loss:

$$SVDloss + ADVloss + CONTloss + ENCODloss. \quad (4)$$

3.2.4 metrics

The model's frame-level performance is analysed using the Area Under the ROC Curve (AUC), after plotting the Receiver Operating Characteristics (ROC) curve. The latter plots the True Positive Rate (TPR) vs the False Positive Rate (FPR) as a function of the detection threshold in the range [0,1], thus summarising the trade-off between TPR and FPR for a predictive model using different probability thresholds. AUC measures the two-dimensional area under the entire ROC curve from (0,0) to (1,1), providing an aggregate measure of performance across all possible detection thresholds which amounts to a sort of probability distribution over the range of thresholds. The AUC thus represents the degree of separability the model can enforce between anomalous and non-anomalous frames. The higher the AUC value, the better the performance.

4. EXPERIMENTS

4.1 Dataset

We validate our approach over several benchmark datasets portraying complex anomalous events in various scenarios involving multiple scenes captured from different angles. All datasets comprise 'normal' video frames for training and a combination of anomalous and non-anomalous frames for testing.

The *CHUK Avenue dataset* contains 16 normal videos for training and 21 videos for testing, for a total of 30,652 frames [20]. Test videos include anomalies like the throwing of objects, walking in the wrong direction, running, and loitering.

The *UCSD anomaly detection dataset* contains surveillance videos of pedestrian walkways [23]. Anomalies include presence of skaters, bikers, small carts and people walking sideways in walkways. The dataset is divided into two parts: Ped1 and Ped2. Ped1 contains 34 normal video samples for training with some perspective distortion and 36 videos samples for testing. Ped2 portrays pedestrians walking parallely to the camera plane, with 16 videos for training and 12 for testing.

4.2 Training

We trained ours Bidirectional Long-Short term memory-based Convolutional Autoencoder Generative Adversarial Network (BILAE- GAN) model on an 8-GPU machine with Quadro RTX 6000 cards having 24 GB VRAM each. Input frames were resized to 128×128 pixels and passed to the BiLAE-GAN architecture. The proposed architecture uses *tanh* activations in the generator and *LeakyRelu* ones in the discriminator. Batch normalisation with *tanh* at the end of each layer helps scaling and adjusting the input features to the interval $[-1,1]$.

The generator uses an Adam optimiser with first-order derivatives. The discriminator uses RMSProp with a 0.00005 learning rate for weight optimisation. In our Generator, each batch of *n* rescaled input frames goes through two layers of depth wise convolution and 4 layers of Bidirectional convolutional LSTM for spatio-temporal feature learning, as shown in Figure 2. Frames are convolved with a kernel of size 5×5 and stride 2 to produce a feature map of size $n \times 64 \times 64 \times 128$. Subsequently, a small kernel of size 3×3 is applied to the respective feature maps to capture spatiotemporal features using a block of Bidirectional convolutional LSTMs. Input frames are encoded to a latent space *Z* of size $n \times 16 \times 16 \times 128$ to be then passed to the decoder for reconstruction.

The decoder uses convolutional 2D transpose layers and batch normalisation to decode the bottleneck features back to the image space ($\bar{X}$). The reconstructed data is remapped to the latent space ($\bar{Z}$) for a consistent comparison between *Z* and $\bar{Z}$. Finally, the generated image $\bar{X}$ and the input image *X* are given as inputs to the discriminator which has the same encoder architecture as the generator, with an additional sigmoid activation for discrimination.Losses are back propagated to the generator for an accurate reconstruction of the input image *X*.

Avneue Dataset

UCSD Dataset

Reconstruction Images

To Train the model our proposed model with BiLSTM take very less time to reconstruct the image within less iterations comparing with LSTM.

Table 1 :Comparision of iterations to Train Model LSTM with BiLSTM

	No. of Iterations	
Data set	BiLSTM	LSTM
UCSD Ped1	3350	45800
UCSD Ped2	6500	60500
CHUK Avenue	4000	75100

4.3 Testing

During testing, Video is divided into frames and tried to find anomaly by loading trained model. By finding the scores of frames. where low score represents anomaly and high score represents normal frames.

To find the Anomaly Score A(X), the square *L*2 distance between input and reconstructed images, rescaled to the interval [0,1]:

sequences_reconstruction_cost=np.array([np.linalg.norm(np.subtract(test_data[i],gan_x[i])) ----------(5)

sa = (sequences_reconstruction_cost - np.min(sequences_reconstruction_cost)) / np.max(sequences_reconstruction_cost)—(6)

sr = 1.0 – sa----------(7)

Threshold taken here to identify anomaly by using statistics.

m=statistics.median(score) (8)

sd=statistics.stdev(score) (9)

threshold=m-sd ----------(10)

Note that we committed to the median of standard deviation of scores of threshold, assess the performance of a model over the whole range of thresholds by measuring the Area Under the ROC Curve (AUC), after plotting the model's Receiver Operating Characteristics (ROC) curve.

4.4 Datasets

We validate our approach over several benchmark datasets portraying complex anomalous events in various scenarios involving multiple scenes captured from different angles. All datasets comprise 'normal' video frames for training and a combination of anomalous and non-anomalous frames for testing. The *CHUK Avenue dataset* contains 16 normal videos for training and 21 videos for testing, for a total of 30,652 frames [5]. Test videos include anomalies like the throwing of objects, walking in the wrong direction, running, and loitering.

The *UCSD anomaly detection dataset* contains surveillance videos of pedestrian walkways [6]. Anomalies include the presence of skaters, bikers, small carts, and people walking sideways in walkways. The dataset is divided into two parts: Ped1 and Ped2. Ped1 contains 34 normal video samples for training with some perspective distortion and 36 video samples for testing. Ped2 portrays pedestrians walking parallel to the camera plane, with 16 videos for training and 12for testing.

4.5 Metrics Used For Evaluation

The model's frame-level performance is analysed using the Area Under the ROC Curve (AUC), after plotting the Receiver Operating Characteristics (ROC) curve. AUC measures the two-dimensional area under the entire ROC curve from (0,0) to (1,1), providing an aggregate measure of performance across all possible detection thresholds which amounts to a sort of probability distribution over the range of thresholds. The AUC thus represents the degree of separability the model can enforce between anomalous and non-anomalous frames. The higher the AUC value, the better the performance. The Equal Error Rate (EER) which corresponds to error rate at which the False Negative Rate equals the False Positive Rate

4.6 Comparison With State Of The Art

The AUC thus represents the degree of separability the model can enforce between normal and anomalous frames. It is worth mentioning that the higher the AUC values, the higher the diagnostic ability of an anomaly detection system. The performance of our BiLAE-GAN on all datasets is compared with that of state-of-the-art *un-supervised* anomaly detection systems in Table 2. The architecture with Convolutional Bidirectional LSTM achieves state-of-the-art results, by a very large margin, with an AUC of 81.9% and EER of 26.5% on CHUK Avenue. On UCSD Ped1 AUC 99.1% and EER 0.84%, UCSD Ped2 99.4% and

EER 5.8%. This is likely due to the relatively short duration (200 frames) of the available training sequences.

Table 2: Comparistion of Metrics with state of the Art

	UCSD Ped1		UCSD Ped2		Avenue	
Unsupervised Methods	AUC	EER	AUC	EER	AUC	EER
MLAD [2019] [8]	82.34	23.50	99.21	2.49	52.82	38.82
ITAE+NFs [2020] [9]	–	–	97.3	–	85.8	–
ROADMAP [2021] [10]	83.4	–	96.3	–	88.3	–
SVD GAN[2021] [4]	73.26	28.75	76.98	23.46	89.82	21.55
Deep Generative Network[2021][3]	85.3	23.6	95.7	12.0	86.9	20.2
DF-ConvLSTM-VAE[2022][2]	88.4	16.7	88.8	12.2	87.2	18.9
DR-STN[2022][1]	98.8	2.9	97.6	6.9	90.8	11.0
BiLAE-GAN Model	99.1	0.84	94.11	5.8	81.9	26.5

Performance comparison in terms of AUC and EER among state-of-art unsupervised anomaly detection architectures, including ours.

In this situation the model tends to reconstruct well the anomalous frames too and falls short when detecting certain anomalies. E.g., the AUC is comparatively low for videos portraying skaters or cyclists in pedestrian pathways these anomalies look closer to normal from the angle (top view) from which the video is captured. This behaviour is to be expected, for Bidirectional LSTM-based models need sufficient sequential data to be excited and perform well.

5. CONCLUSIONS

The proposed Bidirectional LSTM AutoEncoder Generative Adversarial Network (BiLAE-GAN) architecture has a clear edge over state-of-the-art un-supervised anomaly detection methods while using fewer parameters, thanks to using temporal blocks with Bidirectional LSTM for better spatiotemporal feature learning within fewer iterations in training. Our experiments show that our system widely outperforms prior art on the UCSD Ped1 with AUC 99.1,EER 0.84, UCSD Ped2 with AUC 94.11, EER 5.8, and CHUCK Avenue datasets with AUC 81.9,EER 26.5 and can leverage large-scale datasets. The costs of employing ConvBILSTM Lyers are not discussed, nor is it possible to lower them. In the future, the system's accuracy can be further improved by using a 3D feature extractor and we need to validate our architecture on real-time datasets.

REFERENCES:

[1] Y.Cong,J.Yuan,and J.Liu ‘Sparse Reconstruction cost for abnormal event detection,’’ in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 3449–3456.

[2] B. Zhao, L. Fei-Fei, and E. P. Xing, ‘‘Online detection of unusual events in videos via dynamic sparse coding,’’ in *Proc.CVPR*, Jun. 2011, pp. 3313–3320.

[3] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, ‘‘Self-trained deep ordinal regression for end-to-end video anomaly ,’’ 2020, *arXiv:2003.06780*. [Online]. Available: http://arxiv.org/abs/2003.06780

[4] W. Li, V. Mahadevan, and N.Vasconcelos,‘‘Anomaly detection and localization in crowded scenes,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[5] J. K. Dutta and B. Banerjee, ‘‘Online detection of abnormal events using incremental coding length,’’ in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 3755–3761.

[6] J. R. Medel and A. Savakis, ‘‘Anomaly Detection in video using predictive Convolutional long short-term memory networks,’’ 2016, arXiv:1612.00390. [Online]. Available: http://arxiv.org/abs/1612.00390

[7] Y. S. Chong and Y. H. Tay, ‘‘Abnormal event detection in videos using spatiotemporal autoencoder,’’ in Proc. Int. Symp. Neural Netw., vol. 10262. Long Beach, CA, USA: Springer, Dec. 2017, pp. 189–196.

[8] M. Hasan, J. Choi, J. Neumann, A.K.Roy-Chowdhury,andL. S. Davis, ‘‘Learning temporal regularity in video sequences,’’ in Proc. CVPR, Jun. 2016, pp. 733–742.

[9] C. Lu, J. Shi, and J. Jia, ‘‘Abnormal event detection at 150 FPS in MATLAB,’’ in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 2720–2727. [10] W. Luo, W. Liu, and S. Gao, ‘‘Remembering history with

convolutional LSTM for anomaly detection,'' in *Proc. IEEE Int. Conf. Multimedia Expo(ICME)*, Jul. 2017, pp. 439–444.

[11] H. Park, J. Noh, and B. Ham, ''Learning memory-guided normality for anomaly detection,'' in *Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14360–14369.

[12] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, ''Anomalynet: An anomaly detection network for vdeo surveillance,'' *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

[13] M. Mathieu, C. Couprie, and Y. LeCun, ''Deep multi-scale video prediction beyond mean square error,'' in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–14.

[14] W. Liu, W. Luo, D. Lian, and S. Gao, ''Future frame prediction for anomaly detection—A new baseline,'' in *Pro. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[15] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

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

[18] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.

[19] M. Mathieu, C. Couprie, and Y. LeCun,''Deep multi-scale video prediction beyond mean square error,'' in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–14.

[20] W. Liu, W. Luo, D. Lian, and S. Gao, ''Future frame prediction for anomaly detection A new baseline,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[21] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, ''Image-to-image translation with conditional adversarial networks,'' in *Proc. CVPR*, Jul. 2017, pp. 5967–5976

[22] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, ''Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 82–90.

[23] J. Li, X. Mei, D. Prokhorov, and D. Tao, ''Deep neural network for structural prediction and lane detection in traffic scene,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 690–703, Mar. 2017.

[24] A. Stuhlsatz, J. Lippel, and T. Zielke, ''Feature extraction with deep neural networks by a generalized discriminant analysis,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 596–608, Apr. 2012.

[25] Zhiwei Yang , Jing Liu , (Senior Member, Ieee), And Peng Wu "Bidirectional Retrospective Generation Adversarial Network for Anomaly Detection in Videos" in IEEE Access current version August 9, 20 pp. 107842 to 107857.

[26] J. Medel. Anomaly detection using predictive convolutional long short term memory units. 2016.

[27] W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. doi: 10.1109/ICME.2017.8019325.
URL https://doi.ieeecomputersociety.org/10.1109/ICME.2017.8019325.

[28] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems 28: AnnualConference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 802–810, 2015.

[29] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *CoRR*, abs/1511.06309, 2015.
doi: https://doi.org/10.17863/CAM.26485.

[30] Dinesh Jackson Samuel Fabio Cuzzolin Faculty of Technology, Design and Environment Visual Artificial Intelligence Laboratory Oxford Brookes University Oxford, " SVD-GAN for Real-Time Unsupervised Video Anomaly Detection" 15 pages.

[31] Lin Wang,Haishu Tan, Fuqiang Zhou,Wangxia Pengfei Sun"Unsupervised Anomaly Video Detection via a Double Flow ConvLSTM

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Variational Autoencoder" Digital Object Identifier 10.1109 /ACCESS.2022.3165977, VOLUME 10,2022,Pages:44278-44289.

[32] Savath Saypadith,Takao Onoye "An Approach to Detect Anomaly in video using Deep Generative Network"IEEE Access 2021.3126335, Volume 9,2021 pages: 150903-150910

[33] Thittaporn Ganokratanna, Supavadee aramvith,Nicu Sebe "Video anomaly detection using deep residual-spatiotemporal translation network", ELSEVIER-o167-8655/2021 Pattern Recognition Letters 155(2022)143-150

[34] W. Gorr, A. Olligschlaeger, Y. Thompson Assessment of crime forecasting accuracy for deployment of police Int. J. Forecast. (2000), pp. 743-754

[35] C.H. Yu, M.W. Ward, M. Morabito, W. Ding Crime forecasting using data mining techniques 2011 IEEE 11th International Conference on Data Mining Workshops, IEEE (2011, December), pp. 779-786

[36] L.G. Alves, H.V. Ribeiro, F.A. Rodrigues Crime prediction through urban metrics and statistical learning Phys. Stat. Mech. Appl., 505 (2018), pp. 435-443