Video Anomaly Detection Using Self-Attention-Enabled Convolutional Spatiotemporal Autoencoder

Rashmiranjan Nayak
Department of ECE
NIT Rourkela
Odisha-769008, India
rashmiranjan.et@gmail.com

Umesh Chandra Pati
Department of ECE
NIT Rourkela
Odisha-769008, India
ucpati@nitrkl.ac.in

Santos Kumar Das
Department of ECE
NIT Rourkela
Odisha-769008, India
dassk@nitrkl.ac.in

***Abstract*—The process of automatically detecting abnormal video patterns in the intelligent surveillance framework is known as video anomaly detection. However, video anomaly detection is challenging due to inherent research challenges such as equivocal nature, data imbalances, data scarcity, the complex nature of the entities involved in the anomaly, etc. Hence, a self-attention-enabled convolutional spatiotemporal autoencoder is proposed to detect video anomalies efficiently. The proposed Self-Attention-enabled Convolutional Long-Short-Term-Memory Auto-Encoder (SA-ConvLSTM2D-AE)-based video anomaly detector is comprised of three sequential stages: spatial encoder to learn spatial (appearance) features of individual frames, temporal encode-decoder to learn temporal (motion) features of encoded spatial features, and spatial decoder to decode the encoded spatial features for reconstructing the individual frames. Here, the self-attention mechanism is embedded into the convolutional Long Short Term Memory block present in the temporal encoder-decoder section to generate the Spatial-Attention-enabled ConvLSTM block for learning better spatiotemporal features. An efficient threshold selection criteria based on the finding of the optimized Geometric mean value of the sensitivity and specificity from the Receiver Operating Characteristics curve is implemented. The model is trained on only the video frame sequences corresponding to the normal incidents. However, the model poorly reconstructed test frame sequences with video anomalies, as anomalous samples are never exposed during training. Hence, when the anomaly score of individual frames exceeds the selected optimum threshold level, then an anomaly is said to be detected.**

***Index Terms*—Auto-encoders, Convolutional LSTM, Convolutional spatiotemporal autoencoder, Self-attention, Video anomaly detection**

I. Introduction

An abnormal, unusual, or unexpected trend (pattern) present in the video is known as a video anomaly. The automated identification of aberrant video patterns corresponding to anomalous actions or entities in the spatiotemporal dimensions is known as Video Anomaly Detection (VAD) [1]. VAD is highly challenging due to its equivocal nature, inherently class imbalances, high variance within anomalous events, and rareness [2]. Recently, Deep Learning (DL)-based methods using deep reconstruction models, deep predictive models, deep generative models, and deep hybrid models are widely used for VAD due to availability of high end computation facilities, better training strategies, availability of big and consistent video anomaly datasets. In the case of VAD using deep reconstruction models, the video anomaly detectors learn only the normal patterns using the training video data set comprising of video clips involving only normal incidents. Different variants of Auto-Encoders (AEs) have been used in reconstruction models to learn the spatiotemporal patterns of the normal training samples with the help of efficient nonlinear transformations. During testing, frames containing only normal events are reconstructed properly, and frames containing anomalous events are not constructed properly.

Hence, when the anomaly score of any frame exceeds the set threshold, then the frame is treated as an anomalous frame. The basic intuition behind the applicability of the reconstruction model-based VAD is that abnormal events are significantly differs from the normal visual patterns and this visual difference can be quantified in the terms of reconstruction error which is a function of the frame visual statistics [3]. Hence, many deep reconstruction model-based VAD methods such as stacked convolutional AE [4], convolutional Long-Short-Term-Memory (ConvLSTM) AE [5], hybrid spatiotemporal AE [6], R-STAE [7], etc. have been proposed for video anomaly detection. However, there are lots of gray areas related to accuracy and processing time due to ineffective spatiotemporal feature extraction by the AE-based models.

Therefore, in this works, a Self-Attention-enabled Convolutional Long-Short-Term-Memory Auto-Encoder (SA-ConvLSTM2D-AE)-based video anomaly detector is proposed to detect video anomalies accurately. Further, an efficient threshold selection criteria based on Receiver Operating Characteristics (ROC) curve is integrated to the proposed video anomaly detector.

The subsequent sections of the paper are structured in the following manner. The problem formulation and proposed methodology are discussed in Section II and Section III, respectively. Following this, the analysis of the experimental results is presented in Section IV, which is then followed by the conclusion in Section V.

II. Problem Formulation

When there is no direct available information about the positive class, i.e., anomaly cases, the task of anomaly detection is typically treated as an unsupervised learning problem. However, direct information about the negative class, i.e., normal classes or classes with no anomalies, is usually available. Hence, it is intuitive to develop a DL-based VAD model which can learn the underlying patterns of the normal classes to detect the abnormal classes (anomalies). The reconstruction errors, subsequently anomaly scores of the individual frames can be used to quantify the associated spatiotemporal statistics. Further, reconstruction models for VAD are learned by a training dataset, $X_{train} \in R^{N_{train} \times r \times c}$, consisting solely of data samples corresponding to the normal events. The reconstruction model f_θ to detect video anomalies can be learned by minimizing the cost function expressed in terms of reconstruction error as mentioned Eq. 1 for all the normal frame sequences $x_i \in X_{train}$ over all the training sample i.

$$\theta^* = \arg\min_{\theta} \sum_{x_i \in X_{train}} \left\| x_i - f_\theta\left(x_i\right) \right\|^2 \tag{1}$$

Further, the test dataset, $X_{test} \in R^{N_{test} \times r \times c}$, may contain frames corresponding to both normal and abnormal events. Only normal events comply with the learned model and are correctly reconstructed with low reconstruction error during the testing phase. Conversely, anomalies can not be appropriately reconstructed by the same learned model f_θ, resulting in high reconstruction loss. Subsequently, frames with high reconstruction errors (high anomaly scores) above the carefully chosen threshold level will be considered anomalous ones. Hence, the main objective of this work is to develop a reconstruction modeling-based efficient deep AE model to detect video anomalies.

III. Methodology

The proposed DL framework for video anomaly detection relies on a reconstruction modeling approach. This approach is founded on the idea that anomalous events will result in a high anomaly score due to the inability of the trained model to accurately reconstruct the anomalous frames. The proposed Self-Attention-enabled Convolutional Spatiotemporal Auto-Encoder or Self-Attention-enabled Convolutional Long-Short-Term-Memory Auto-Encoder (SA-ConvLSTM-AE) to detect video anomalies is depicted in Fig. 1. The intuition behind the use of with Self-Attention Convolutional Long-Short-Term-Memory (SA-ConvLSTM) cell is to extract the efficient spatiotemporal features by applying attention in the spatiotemporal domain during the autoencoder training in an end-to-end pipeline. The methodologies of the proposed framework can be explained in two key steps: model development and anomaly detection.

A. Model Development

The proposed SA-ConvLSTM-AE model to detect video anomalies is inspired by temporal regularity learning [3], the use of spatiotemporal autoencoders for anomalous event detection [4], and the effectiveness of self-attention-enabled Convolutional LSTM for spatiotemporal prediction [8]. The proposed model, as presented in Fig. 1, comprises three sequential stages: spatial encoder to learn spatial (appearance) features of individual frames, temporal encode-decoder to learn temporal (motion) features of encoded spatial features, and spatial decoder to decode the encoded spatial features for reconstructing the individual frames. The numbers mentioned as quadruplets in Fig. 1 are the output dimensions of each layers represented in the form of "$F_{TW} \times Height\ of\ the\ frame \times Width\ of\ the\ frame \times Number\ of\ channels$." Similarly, the numbers mentioned as triplets in Fig. 1 are model parameters corresponding to the operation carried in the particular layer represented in the form of "$Number\ of\ filter,\ Filter\ Kernel\ size,\ Stride$." The number of frames in the input frame sequence block or the length of the temporal sliding window, i.e., F_{TW}, is a vital model parameter that significantly affects model performances. Usually, the model can learn more discriminative regularity scores with higher values of F_{TW} at the cost of training time and computational resources [3]. Conversely, very low values of F_{TW} severely degrade the model performance due to poor learning of the temporal regularity among the frame sequences. Experimentally, it is noted that $F_{TW} = 10$ provides an optimum trade-off between the training time and discriminative ability of the model for the given experimental setup. Usually, video anomaly detection is a coarse level understanding of the scene with spatiotemporal context to flag the anomalies [9]. Hence, the proposed model uses two dimensional convolution and LSTM operations to develop a computationally efficient model for video anomaly detection in real-world scenario.

1) Spatial Encoder: The spatial encoder is realized using three convolutional layers (Conv2D layers) to extract spatial features efficiently as convolution operation preserves the inter-spatial relationship of the pixels in the frame [4]. The output of the each Conv2D layer is normalized using layer normalization to smooth the training process. The number of filters (32, 64, and 128) of each Conv2D layer in the spatial encoders is doubled in the direction of increase in depth to extract more complex features effectively. Similarly, the kernel size of the filters (9×9, 7×7, and 5×5) in the Conv2D layer is in decreasing order in the direction of increase in depth to extract global to local features effectively. The individual input frames are encoded by the spatial encoder with extracted spatial features in each convolutional layer sequentially. A spatial encoded feature vector is created by concatenating the encoded features of consecutive F_{TW} frames and subsequently, inputted to the temporal encoder-decoder.

2) Temporal Encoder Decoder: The temporal encoder-decoder is realized using three Self-Attention-enabled Convolutional LSTM (SA-ConvLSTM2D) layers. The middle SA-ConvLSTM2D layer acts as a bottleneck of this deep autoencoder where the spatiotemporally encoded frame is generated. It helps in increasing the performance of the SA-ConvLSTM-AE by removing the redundant spatiotemporal features and best fitting the salient spatiotemporal features in the available latent feature space. Each SA-ConvLSTM2D layer is developed using a Self-Attention-enabled ConvLSTM (SA-ConvLSTM) blocks, as mentioned in [8]. Convolutional LSTM (ConvLSTM) cell [10] uses convolution operation instead of matrix operation in contrast to a fully connected LSTM block to extract better spatiotemporal features with fewer model parameters when used at the input-to-hidden or hidden-to-hidden links by exploiting the strength of both CNN (efficient in spatial feature learning) and LSTM (efficient in temporal feature learning) [11]. However, the prediction or reconstruction of the current frame using spatiotemporal modeling can be significantly improved without increasing the computational complexity significantly if ConvLSTM operation is performed only for most relevant past features. Hence, Self-attention [12] is embedded into the ConvLSTM cell to develop SA-ConvLSTM cell as shown in Fig. 2 [8]. Self-attention (intra-attention) is a special type of attention mechanism that captures the most relevant features of the input sequence by exploiting the relationship among the

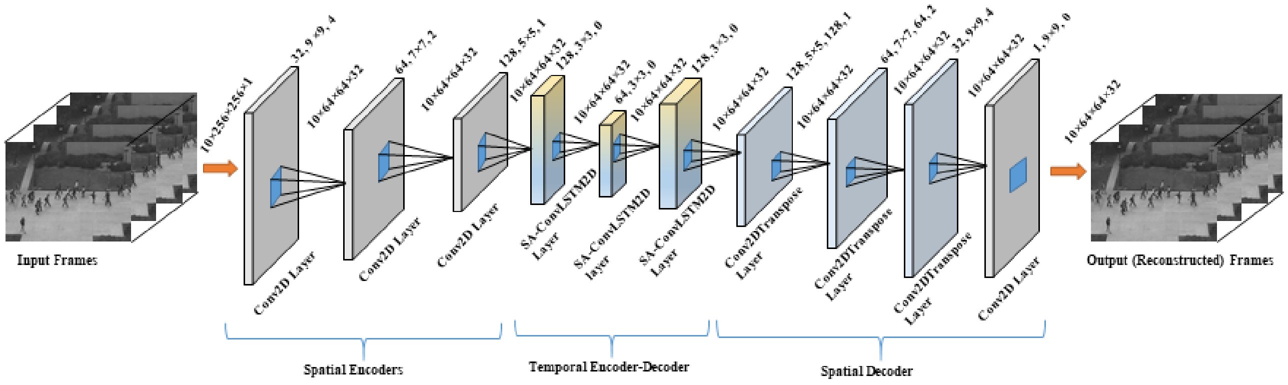

Fig. 1. Architecture of the proposed Self-Attention-enabled ConvLSTM Auto-Encoder (SA-ConvLSTM-AE).

different elements of the same sequence only. It is efficient and more interpretable for extracting vital spatiotemporal features [12]. The core difference between ConvLSTM cell and SA-ConvLSTM cell is the embedding of the Self-Attention Memory (SAM) module M in to ConvLSTM cell to generate SA-ConvLSTM cell. The end-to-end pipeline of the SAM module can be segregated into three parts: feature aggregation for obtaining the global context information, the memory updation, and generation of the output [8]. The hidden state in ConvLSTM at time step t is represented by H_t in the self-attention module. Q_h is the query, K_h is the key, V_h denotes the value based on the 1×1 convolution on the feature, and H_t is the output. In the SAM, the aggregated feature Z_h is created by applying self-attention to H_t and a different feature Z_m, where Z_m is calculated by querying on K_m and visiting to V_m. Further, K_m and V_m are both mappings of the memory M_{t-1}. The fusion of the Z_h and Z_m are performed to form Z. The memory is then updated using a gating mechanism using Z and the original input H_t. The final result is a dot product between the value of the output gate $o_t^{'}$ and the updated memory M_t [8].

3) Spatial Decoder: The spatial decoder is realized using three convolutional transpose layers (Conv2DTranspose layers) and one final Conv2D layer to reconstruct the individual frames from the encoded sequences inputted by the temporal decoder. The output of each Conv2DTranspose layer is normalized using layer normalization to smooth the training process and reconstruction. The number of filters (128, 64, and 32) of each Conv2DTranspose layer in the spatial decoders is selected exactly in the reverse order of the spatial encoder in the direction of increase in depth to inverse the effects of convolution using the deconvolution process efficiently. Similarly, the kernel size of the filters (5×5, 7×7, and 9×9,) in the Conv2DTranspose layers are in increasing order in the direction of increase in depth to recover the embedded features in the feature vectors, efficiently. Finally, an additional Conv2D layer with a Sigmoid activation function is used to transform the reconstructed frames to the appropriate resolution that is the same as that of the input frames.

Training a deeper RNN network is sometimes become difficult due to gradient vanishing problem. Hence, Rectified Linear Unit (ReLU) is used as the activation functions in each Conv2D layers, SA-ConvLSTM Layers, and Conv2DTranspose layers, except the final layers, to minimize the likelihood of the gradient to vanish. The stride lengths in different layers are chosen carefully to reduce the computational complexity by decreasing the overlapping of the receptive fields without compromising on the network efficiency. The order of the increase and decrease of the stride length in different layers of the SA-ConvLSTM-AE are selected in proper order to maintain the dimensional consistency. The padding is set to “SAME” during training of the model to ensure the sizes of the input and out frames of the convolution or deconvolution operations remain the same.

B. Anomaly Detection

The proposed SA-ConvLSTM-AE model can be used to calculate the anomaly scores $A_{Score}\left(t\right)$ of the individual frames. If $A_{Score}\left(t\right)$ exceeds the selected threshold A_{Th}, then the corresponding frame is considered as anomalous one. Hence, calculation of $A_{Score}\left(t\right)$ and finding optimum A_{Th} are two crucial steps in VAD.

1) Anomaly Score: The $A_{Score}\left(t\right)$ is calculated from the visual statistics of the reconstructed frames as follows. The pixel reconstruction error $e_{pixel}\left(p,q,t\right)$ at spatial location (p,q) of a frame at time t for the given intensity level I can be calculated using Eq. 2 [3].

$$e_{pixel}(p,q,t) = \left\|I(p,q,t) - \hat{I}(p,q,t)\right\|_2 \tag{2}$$

Here, the reconstructed frame $\hat{I}(p,q,t)$ is produced by the learned model by model SA-ConvLSTM AE, i.e., $F_{SA-ConvLSTM-AE}$ as expressed in Eq. 3.

$$\hat{I}(p,q,t) = F_{SA-ConvLSTM-AE}\left[I(p,q,t)\right] \tag{3}$$

Then, frame-level reconstruction error of the individual frames at time t may be calculated by using Eq. 4.

$$e_{frame}\left(t\right) = \sum_{(p,q)} e_{pixel}\left(p,q,t\right) \tag{4}$$

Then, the anomaly score $A_{Score}\left(t\right)$ in the range of 0 to 1 can be calculated using

$$A_{Score}\left(t\right) = \frac{e_{frame}\left(t\right) - e_{frame t_{\min}}\left(t\right)}{e_{frame t_{\max}}\left(t\right)} \tag{5}$$

2) Thresholding: Generally, the datasets used for the video anomaly detection are inherently sufficiently imbalanced ones and hence, VAD is an imbalanced classification problem. Therefore, threshold for detecting anomalies, i.e., A_{Th}, must be set optimally to achieve best detection accuracy with desired sensitivity. Hence, finding optimum threshold is one of the crucial step in the VAD. ROC curve is a diagnostic

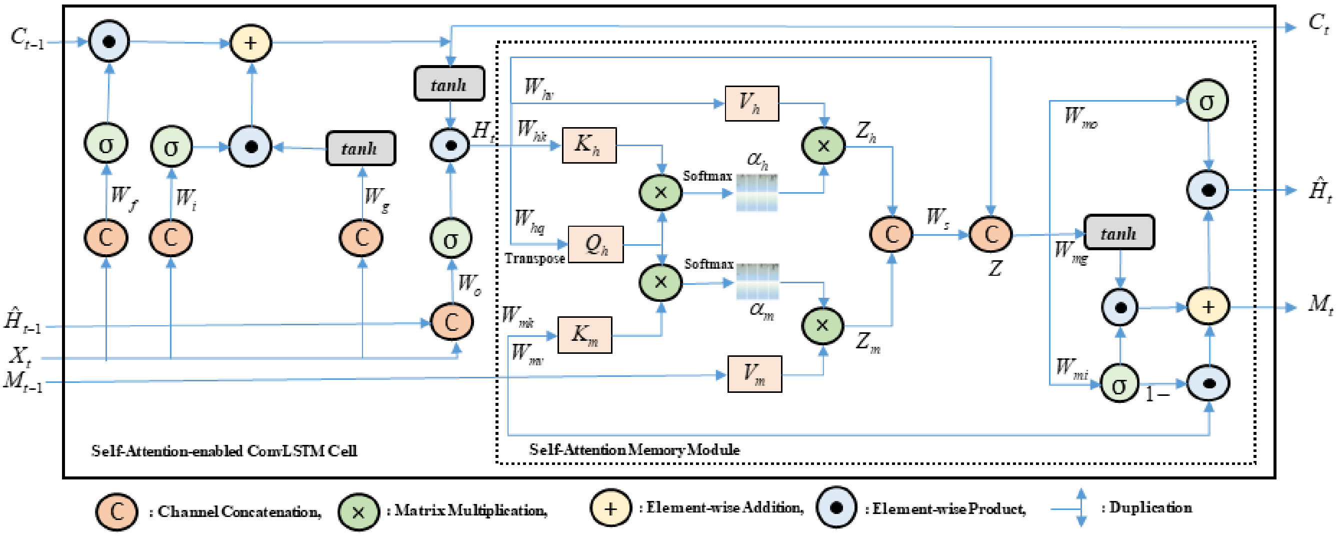

Fig. 2. Self-Attention-enabled ConvLSTM (SA-ConvLSTM) cell.

graph used to the performance of a classification model at all classification thresholds for a given test dataset [13]. Hence, an optimal threshold can be localized on the ROC curve for achieving optimal balance between False Positive Rate (FPR) and True Positive Rate (TPR) by optimizing the Geometric mean (G-Mean) of Sensitivity and Specificity as expressed in Eq. 6 [14]. In first step, G-Mean is calculated for each threshold levels of the ROC curve. In second step, location of the index for the largest G-Mean score is used as optimum threshold value to detect the frame level video anomalies.

$$\begin{aligned} G_{Mean} &= \sqrt{\{Sensitivity \times \ Specificity\}} \\ \Rightarrow G_{Mean} &= \sqrt{\{TPR \times (1 - FPR\}} \end{aligned} \tag{6}$$

IV. Experimental Results Analysis

The implementation details of the experiment being performed to develop the proposed model and the results are discussed in this section.

A. Experimental Setup

The experimental configuration employed for developing the proposed model consists of a high-performance graphical computing system, which includes an Intel Xeon Silver 4214 processor with 64-bit architecture and 12 cores, an NVIDIA RTX A4000 graphics card with 16 GB of memory, and 64 GB of DDR4 RAM. The system is equipped with the Ubuntu 20.04 LTS operating system, Tensorflow 2.6 machine learning framework, Python 3.8 programming language, CUDA 11.2 parallel computing platform, and CuDNN 8.1 deep neural network library.

B. Datasets

A static or stationary surveillance camera is used to acquire the UCSD Pedestrian 2 (Ped2) dataset [15] at an fps of 10 from an elevation covering an outdoor scene, namely the pedestrian walkways. Ped2 has 28 total video clips, i.e., the number of training and testing video clips is 16 and 12, respectively. The training clips consist of only normal events, i.e., only pedestrians. However, the testing clips consist of both normal events and video anomalies. Ped2 contains 4560 frames with 12 video anomalies. The video anomalies are caused due to the circulation of the non-pedestrian objects and abnormal motion patterns of the pedestrians in the walkways. Hence, all other objects such as cars, bikers, skaters, and vehicles are treated as video anomalies apart from pedestrians. Illumination variation, variable crowd density ranging from sparse to very crowded, scale changing of the objects, and perspective distortion are the essential research challenges provided by the Ped2 dataset.

C. Prepossessing

The resolutions of the input frame sequences (Ped2) are changed to a fixed resolution of 256 × 256 to match the input dimensions of the developed model. Frames are converted into gray scale images and normalized in the range of 0 to 1. To increase the volume of the training dateset, temporal data augmentation technique [3] is applied by concatenating three different strides of ten consecutive frames taken from the train video sequences for different stride values of 1, 2, and 3, sequentially.

.

D. Model Training

The model is trained using Adam optimizer with MSE loss function and optimization parameters such as learning rate (l_r) of 0.001, decay of 0.00001, ϵ of 0.000001 for epoch of 250 with batch size of 4, patience of 50, and F_{TW} of 10. The model loss during training time for the base line models: only spatial, ConvLSTM-AE (spatial + temporal) and SA-ConvLSTM-AE (spatial + temporal + attention) with F_{TW} =10 are shown in Fig. 3. It is evident that models are properly trained due to converge of the loss curves towards loss value of zero. Further, addition of self-attention helps in completion of the training faster.

E. Performance Evaluation

The performance of the proposed video anomaly detector can be evaluated in terms of the quantitative analysis and qualitative analysis as follows.

1) Quantitative Analysis: The quantitative performances are evaluated using both Area Under Curve (AUC) and Equal Error Rate (EER) that are evaluted form the ROC curve. For anomaly detection applications, higher AUC and lower EER are preferred. The ROC curve for the ConvLSTM-AE model

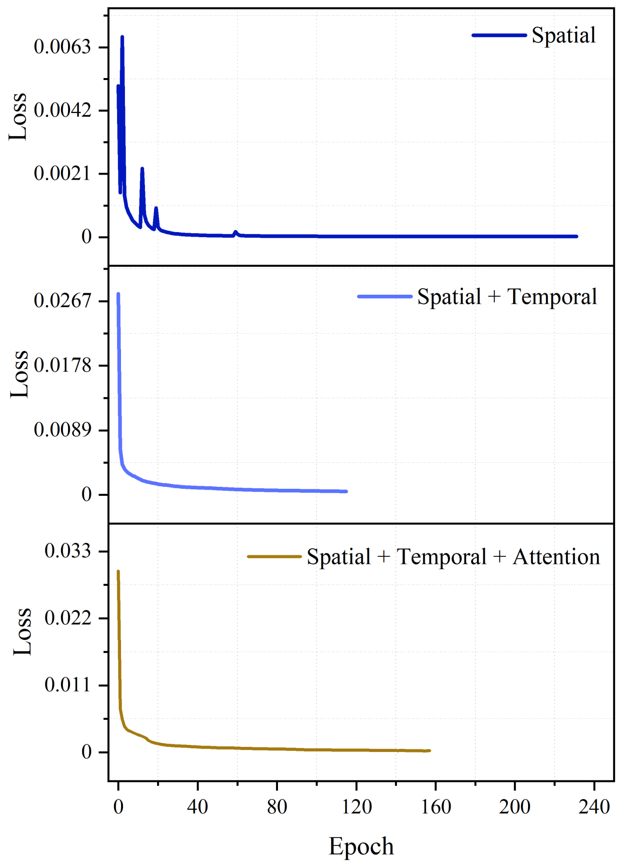

Fig. 3. Model loss during training of the models for UCSD Ped2 dataset..

for Ped2 dataset is presented in Fig. 4. The AUC and EER for the ConvLSTM AE-based video anomaly detector for the Ped2 dataset is 0.855 and 0.2209, respectively. Optimal threshold value, i.e., $A_{Th}^{Optimal}$, is found to be 0.064734 with largest G-Mean value of 0.786. Further, the ROC curve for the proposed model, i.e., SA-ConvLSTM-AE model - based video anomaly detector for Ped2 dataset is presented in Fig. 5. Here, the AUC and EER for the proposed video anomaly detector for the Ped2 dataset is 0.926 and 0.1316, respectively. Optimal best threshold value is found to be 0.075449 with largest G-Mean value of 0.873. Form this quantitative analysis, it is clearly evident that the embedding of self-attention into the ConvLSTM cell significantly helps in better spatiotemporal feature learning.

2) Qualitative Analysis: The qualitative performances can be evaluated using the visualization results of anomaly score for different test sequences as represented in Fig. 6 and Fig. 7. Form these figures, it is evident that anomaly score increases beyond the threshold level, when a person riding cycle passes the pedestrian path. This happens due to significant appearance as well as motion pattern deviation causes by the bicycle rider circulation. Subsequently, the corresponding frames have been marked as anomalous ones by the proposed video anomaly detector.

F. Ablation Study

A systematic study is performed to investigate the importance of the individual blocks of the proposed SA-ConvLSTM-AE-based video anomaly detector as presented in Table I. It is evident that incorporating self-attention to the ConvLSTM cell increases the overall performance of the proposed framework.

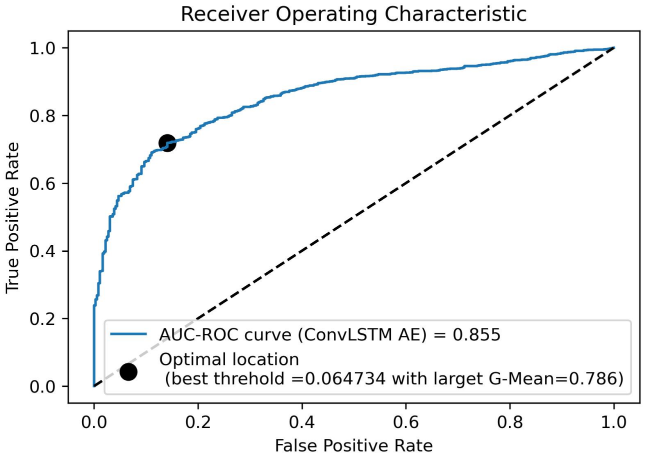

Fig. 4. ROC curve of the ConvLSTM AE for UCSD Ped2 dataset.

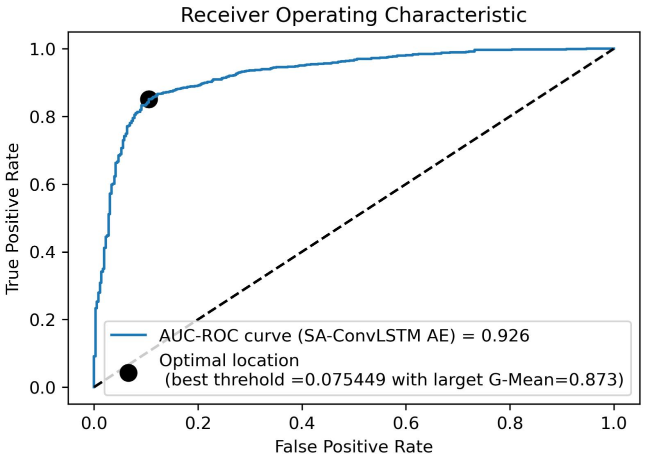

Fig. 5. ROC curve of the SA-ConvLSTM-AE model for Ped2 dataset.

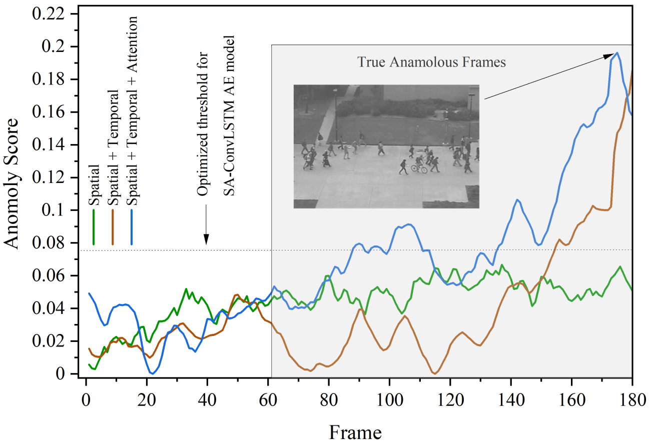

Fig. 6. Visualization results of $A_{Score}\,(t)$ for Test001 clip of Ped2 dataset.

G. Comparative Analysis

A comparative analysis of the proposed video anomaly detector with the SOTA for the UCSD Ped2 dataset is presented in Table II, and the results are found to be quite promising. Further, from this table, the accuracy of the previous methods may be due to the ineffective spatiotemporal

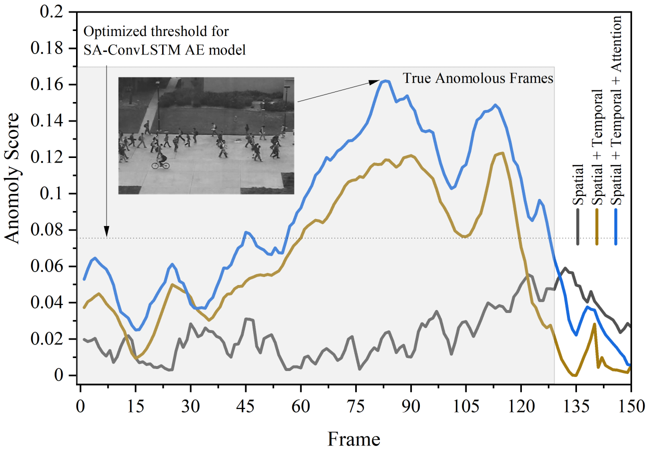

Fig. 7. Visualization results of $A_{Score}\ (t)$ for Test005 clip of Ped2 dataset.

TABLE I
PERFORMANCE EVALUATION WITH RESPECT TO DIFFERENT BLOCKS

Model Blocks		Performance Metrics			
Conv-LSTM2D Layers	**SA-Conv-LSTM2D Layers**	**AUC (%)**	**EER (%)**	**No. of Parameters**	**FPS**
✓	✗	85.54	22.09	3797249	12.85
✓	✓	**92.65**	**13.16**	3822849	**12**

feature extraction from the input videos by the proposed DL models. However, the proposed model attempts to extract better spatiotemporal features from the input videos by adding the self-attention mechanism into the ConvLSTM blocks.

TABLE II
COMPARATIVE ANALYSIS OF THE PROPOSED MODEL

Ref.	**Year**	**Method**	**AUC (%)**
[3]	2016	ConvAE	90.0
[16]	2017	ConvLSTM-AE	88.1
[4]	2017	STAE	88.9
[6]	2020	SiTGRU	86.2
[17]	2020	Pang et. al	83.2
[7]	2021	R-STAE	83
[18]	2022	Guo et. al	88.1
Proposed Work	**2023**	**SA-ConvLSTM AE**	**92.65**

V. CONCLUSION

An efficient SA-ConvLSTM-AE-based video anomaly detector is proposed. The complete end-to-end pipeline of the proposed model is optimized to improve the model performance. The improved performances of the proposed framework is successfully validated with one of the challenging bench-marked video anomaly detection datasets. The model performance is increased by at least by 8% (increase in AUC from 85.5% to 92.6%) by embedding self-attention into the convolutional LSTM-based autoenoder. In future, the proposed model can be investigated with other bench-marked video anomaly datasets to validate its generalization ability.

REFERENCES

[1] B. Yang, J. Cao, R. Ni, and L. Zou, “Anomaly detection in moving crowds through spatiotemporal autoencoding and additional attention,” *Advances in Multimedia*, vol. 2018, 2018.

[2] R. Nayak, U. C. Pati, and S. K. Das, “A comprehensive review on deep learning-based methods for video anomaly detection,” *Image and Vision Computing*, vol. 106, p. 104078, 2021.

[3] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, “Learning temporal regularity in video sequences,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.

[4] Y. S. Chong and Y. H. Tay, “Abnormal event detection in videos using spatiotemporal autoencoder,” in *Int. Symp. on Neural Networks*. Springer, 2017, pp. 189–196.

[5] S. Sudhakaran and O. Lanz, “Learning to detect violent videos using convolutional long short-term memory,” in *Proc. IEEE 14th Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–6.

[6] H. Fanta, Z. Shao, and L. Ma, “Sitgru: single-tunnelled gated recurrent unit for abnormality detection,” *Information Sciences*, vol. 524, pp. 15–32, 2020.

[7] K. Deepak, S. Chandrakala, and C. K. Mohan, “Residual spatiotemporal autoencoder for unsupervised video anomaly detection,” *Signal, Image and Video Processing*, pp. 1–8, 2020.

[8] Z. Lin, M. Li, Z. Zheng, Y. Cheng, and C. Yuan, “Self-attention convlstm for spatiotemporal prediction,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 531–11 538.

[9] W. Sultani, C. Chen, and M. Shah, “Real-world anomaly detection in surveillance videos,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6479–6488.

[10] V. Patraucean, A. Handa, and R. Cipolla, “Spatio-temporal video autoencoder with differentiable memory,” *arXiv preprint arXiv:1511.06309*, 2015.

[11] R. Nayak, U. C. Pati, and S. Kumar Das, “Video anomaly detection using convolutional spatiotemporal autoencoder,” in *Proc. IEEE Int. Conf. on Contemporary Computing and Applications (IC3A)*, 2020, pp. 175–180.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[13] F. Provost, “Machine learning from imbalanced data sets 101,” in *Proc. AAAI’2000 workshop on imbalanced data sets*, vol. 68, no. 2000. AAAI Press, 2000, pp. 1–3.

[14] J. Brownlee, “A gentle introduction to threshold-moving for imbalanced classification,” https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/, Jan. 2021.

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, “Anomaly detection in crowded scenes,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1975–1981.

[16] W. Luo, W. Liu, and S. Gao, “Remembering history with convolutional lstm for anomaly detection,” in *IEEE Int. Conf. on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 439–444.

[17] G. Pang, C. Yan, C. Shen, A. v. d. Hengel, and X. Bai, “Self-trained deep ordinal regression for end-to-end video anomaly detection,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020, pp. 12 173–12 182.

[18] A. Guo, L. Guo, R. Zhang, Y. Wang, and S. Gao, “Self-trained prediction model and novel anomaly score mechanism for video anomaly detection,” *Image and Vision Computing*, vol. 119, p. 104391, 2022.