Neural Processing Letters (2021) 53:4677–4692
https://doi.org/10.1007/s11063-021-10618-3

Residual Spatiotemporal Autoencoder with Skip Connected and Memory Guided Network for Detecting Video Anomalies

S. Chandrakala[1] · V. Srinivas[1] · K. Deepak[2]

Accepted: 2 August 2021 / Published online: 25 August 2021

Abstract
Real-time video anomaly detection and localization still prevail as a challenging task. Autoencoders are expected to give high reconstruction error for abnormal events than normal events while trained on video segments of normal events. Nevertheless, this assumption is not always true in practice. Sometimes the autoencoder offers better generalization. Therefore, it also reconstructs abnormal events well, leading to slightly degraded performance for anomaly detection. To alleviate this issue, we propose a Skip connected and Memory Guided Network (SMGNet) for video anomaly detection. The memory guided network with skip connection help in avoiding loss of meaningful information such as foreground patterns, in addition to memorizing significant normality patterns. The effect of augmenting memory guided network with skip connection in the residual spatiotemporal autoencoder (R-STAE) architecture is evaluated. The proposed technique achieved improved results over three benchmark datasets.

Keywords Video anomaly detection · Normality modeling · Memory guided network · Spatio temporal autoencoders · Residual blocks

1 Introduction

The task of anomaly detection has recently gained a lot of attention in the field of video surveillance. Video anomaly detection (VAD) and localization play an inevitable role in ensuring public and private safety. Unlike the supervised video classification task, there exists various challenges that one faces when building a VAD system. One of the challenges being insufficient training data for anomalous activities, which creates an imbalance between normal and anomalous samples. The task becomes incredibly challenging since the data points

✉ S. Chandrakala
chandrakal@cse.sastra.edu; sckala@cse.iitm.ac.in

V. Srinivas
srinivasvasudevan2000@gmail.com

K. Deepak
deepakk2@srmist.edu.in

[1] Intelligent Systems Group, SASTRA Deemed to be University, Thanjavur, India

[2] SRM Institute of Science and Technology, Ramapuram, Chennai, India

lie in a higher dimension. Also, there are varying kinds of anomalies based on the scenario. For example, running in the middle of the road might be considered anomalous while running in a park is not. Due to these challenges, it becomes impractical to address VAD with typical supervised video event classification techniques. Conventionally, it is treated as an outlier detection problem. A normality model is trained based on normal activities present in the training data, and deviations from the normality model are detected as anomalies while testing.

Recent techniques proposed for VAD are based on unsupervised deep learning architectures, which involve training an autoencoder based on normal video events, and the anomalous activities are then identified based on the reconstruction error. However, few of these methods [14,43] solely depend on a 2D convolutional autoencoder (2D-CAE) or fully-connected autoencoder (FC-AE) in which the convolution and pooling operations are performed only in the spatial dimensions, in turn, fails to capture the temporal characteristics of abnormal activities, which are essential for video anomaly detection. To alleviate this issue, few approaches [3,44] incorporated 3D-convolution layers and convolutional LSTM (C-LSTM) layers to autoencoder to derive motion information from video events.

In [7], we proposed a Residual Spatio-temporal Autoencoder (R-STAE) approach for normality modeling. Spatio-temporal features are extracted from video segments and residual blocks are used to mitigate the vanishing gradients problem. This approach provides incremental performance consistently for three datasets used for abnormality detection. The problem with normality model-based approaches is that autoencoders tend to generalize well so that few anomalous activities might also be reconstructed well. To avoid this issue, the memory-guided network is used to capture and store the significant normal patterns in MemAE [13] approach. Inspired by this approach, we propose a Skip-Connected and Memory Guided Network (SMGNet) as an extension to our R-STAE [7] based approach. Unlike the MemAE [13] approach where the memory module is augmented in the convolution autoencoder (CAE), we propose the skip connected memory module in the R-STAE architecture instead of CAE to improve detection performance. The proposed architecture captures significant normal patterns for normality modeling. Memorizing the significant normal patterns sometimes leads to loss of information while reconstructing normal foreground objects since only a minimal set of significant normal patterns are used while reconstruction. To overcome this issue, a skip connection is also introduced in the SMGNet approach to compensate for this kind of loss of information. The proposed SMGNet approach is capable of performing better than the state-of-the-art models.

2 Related Work

So far in the literature, the techniques proposed for VAD fall under the following categories: (1) Modeling events using hand-crafted feature based techniques which make use of features such as histogram of gradients [4], histogram of optical flow [5], trajectories [36], 3D-gradients [19], etc. Extracting hand-crafted features is time-consuming, and also their representation capabilities are limited for complex visual interactions. (2) Unsupervised deep learning-based methods which involve training an autoencoder based on normal video events and the anomalous activities are then identified based on the reconstruction error. In this section, a few important normlaity modeling methods are discussed for VAD.

Feng et al. [10] propose to use PCANet [2] modeled using spatiotemporal gradients of normal image patches to derive the deep features. Then to train a generative model for these normal patterns, the Deep GMMs [33] are used. The likelihood scores given the deep GMM for the testing patterns are used as anomaly scores to detect abnormal activities. Srivastava et al. [37] proposed a composite FC-LSTM model that merges an autoencoder and predictive LSTM model. Basically, the autoencoder sometimes learns insignificant features of the input data by memorization. But the memorization of input patterns does not help much in probabilistically predicting the future frames. Consequently, the role of a future frame predictor is to incorporate the memory of the previous frames. But it does not cope well with the generalized loss function of the autoencoder. But the composite LSTM [10] model alleviates these issues in forming more significant video representation to predict the future frames.

In [31], the C-LSTM model was used along with a composite LSTM model that follows an encoder-decoder architecture. Interestingly, this architecture consists of two streams, one is used for reconstruction, and another is used for prediction. In [6], a recurrent autoencoder model is combined with an LSTM to learn the temporal features from input image patches to detect video forgery.

Hinami et al. [15] proposed a novel approach for recounting the anomalous events as they are detected. In this approach, firstly they train a Fast-RCNN model [12] on the large-scale Visual Genome [20] and COCO [25] datasets to detect the activities and objects. The frame-wise features are extracted from the last fully connected layer, and anomalies are detected using a one-class SVM. Alternatively, the likelihood score is also obtained with respect to kernel density estimate with Radial Basis Function (RBF) kernel for further decision making.

The 2D-ConvNets are highly effective in learning representations for image classification, but they are unable to capture the temporal changes present in consecutive frames to solve video related problems. For this purpose, the 3D-Convolution architectures used for action recognition [39] are used to design the 3D-autoencoders to obtain meaningful representations that are invariant to intra-class spatiotemporal changes [45]. This approach uses stacked frames as an input to the 3D-filters as done in Fully connected AE [14] approach. The feature maps obtained out of 3-D filters are used to model the spatiotemporal changes. The prediction stream better handles the issue of poorly reconstructed normal events by the autoencoder stream. Local temporal coherence was taken into consideration while designing the prediction loss.

Sun et al. [38] proposed a normality model by exploring the Growing Neural Gas (GNG) [11] algorithm with Spatio-temporal interest point features as inputs extracted from video snippets. They incorporate online updates in GNG using techniques such as neuron deletion, insertion, early stopping criteria, and imposing adaptive learning rates. During the testing phase, the patterns that are far away from the nearest neighbors in the trained model are considered anomalies.

As an extension to normality clustering-based approach [17], Ionescu et al. extended anomaly detection as a binary classification problem [16]. Initially, an unsupervised feature learning framework was proposed with the help of object-centric autoencoders to learn the motion and appearance based features. Secondly, the training data is partitioned into clusters of normal patterns. Then they use a one-vs-rest approach by treating one of the clusters as normal, and the rest acts as dummy anomalies. During testing, a video patch is labeled as abnormal if the binary classifier provides a negative score for the patch.

Ramachandran et al. [34] explore the Siamese neural network to develop the nearest neighbour scheme as an alternative to the hand-crafted feature-based representations. They model a Siamese neural network to classify between normal and anomalous video patches by using similarity measures. Firstly, an exemplar model comprised of unique normal patterns is built using training data of normal events only. The anomaly scores for the test video patches are assigned based on the nearest neighbour scoring between the new testing patches and the exemplar model learned beforehand.

Li et al. [23] proposed a Multivariate Gaussian Fully Convolution Adversarial Autoencoder (MGFC-AAE) for anomaly detection and localization in videos. Their approach works based on the fact that the latent representations of normal video segments will be under a prior distribution obtained out of the trained autoencoder. Whereas the anomalous videos do not fall under this distribution. To derive the latent representations, CNN layers are used in the encoder part of the network. An energy-based technique is utilized to get the anomaly score of a video segment based on the probability score obtained out of the trained model. Employing a two-stream network with gradients and optical flow as inputs proved to be effective in attaining meaningful representations of the video segments, which in turn comprehensively improves the detection results. Finally, a multi-scale patch-based structure is also employed to handle the varying perspective of a few scenes.

A semi-supervised learning approach for VAD using dual discriminator based GAN architecture is proposed in [9]. Unlike the other techniques, this approach focuses more on representing the motion representation. During training, the future frames are predicted through the generator, and they try to coerce the predicted frames to be similar to the ground truth. Both the frame and motion discriminators are utilized to force the generator to construct much realistic successive frames. The role of the frame discriminator is to evaluate whether the upcoming frames are real. The purpose of the motion discriminator is also the same with optical flows as inputs. The generated sequence of frames is used to estimate the fake optical flow fields. During testing, the predicted frames are evaluated based on a regularity score. By intuition, the frames providing low regularity scores are detected as abnormal frames.

3 Skip Connected and Memory Guided Network (SMGNet) for Video Anomaly Detection

Recent approaches pose detecting video anomalies as an outlier detection problem, where the focus is on modeling the patterns of normal events, and the events that deviate from the normality model are treated as anomalies. The existing autoencoder architectures use 3D-Convolution layers and LSTM layers to effectively capture the spatiotemporal information present in the videos. In a recent work [24], a two-stream autoencoder architecture is used to extract appearance and motion information, respectively.

A notable issue with the conventional autoencoder models is that there is always a possibility for the autoencoder to generalize well, even for anomalous frames, thereby reducing the reconstruction error, which is unfavorable. To alleviate this issue, a memory module was used to capture and store the prototypical normal patterns in MemAE [13] approach. Inspired by this approach, we propose Skip connected and Memory Guided Netowork (SMGNet) as an extension to our R-STAE [7] based approach. Unlike the MemAE [13] approach where the memory module is augmented in the convolution autoencoder (CAE), we propose the skip connected memory module in the R-STAE architecture instead of CAE to improve detection performance. The proposed architecture memorizes significant normal patterns for reconstruction based normality modeling. Memorizing the significant normal patterns sometimes leads to loss of information while reconstructing normal foreground objects since only a minimal set of prototypical normal patterns are used while reconstruction. To overcome this issue, a skip connection is also introduced in the SMGNet approach to compensate for this kind of loss of information.

3.1 Normality Modelling Using SMGNet

The architecture of the proposed SMGNet is shown in Fig. 1. The encoder consists of two 3-D convolution layers one Convolution-LSTM (C-LSTM) layer. The output channels of the 3D-convolution layers are fixed as 128, and 64 units respectively. Simple LSTMs are not able to hold on to appearance information of video sequences. To address this issue, C-LSTM was introduced where all the states are 3D tensors and can accommodate spatial dimensions. Let x_t be the value of input sequence at time step t, and hidden state is given by h. The gates are denoted as i, f, o and the cell output is given by C. The convolution operator is given by $\star$, $\circledcirc$ is the Hadamard product, W denotes the weight matrices and bias vectors are given by b. As mentioned in [42], Conv.LSTM is given by:

$$i_t = \sigma(W_i \star [x_t, h_{t-1}] + W_i \circledcirc C_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_f \star [x_t, h_{t-1}] + W_f \circledcirc C_{t-1} + b_f) \tag{2}$$

$$o_t = \sigma(W_o \star [x_t, h_{t-1}] + W_o \circledcirc C_t + b_o) \tag{3}$$

$$C_t = f_t \circledcirc C_{t-1} + i_t \circledcirc \tanh(W_c \star [x_t, h_{t-1}] + b_c) \tag{4}$$

$$h_t = o_t \circledcirc \tanh(C_t) \tag{5}$$

Residual Networks: The residual blocks used in the proposed SMGNet architecture is presented in Fig. 1, and the architecture configuration is presented in Table 1. The residual network makes use of a skip connection apart from the existing layers. This helps in avoiding the loss of meaningful information from the previous convolution layers and also bestow for gradient flow while backpropagation, thus helps in taking control over the vanishing gradients. The equation of a residual block with input x is given by,

$$\mathbf{y_r} = F(x) + x \tag{6}$$

Here, x denotes encoded feature maps before passing them into the residual block. $F(x)$ refers to encoded feature maps obtained from the residual blocks, and $\mathbf{y_r}$ denotes the encoded representation obtained by adding x and $F(x)$. ReLU activation function is used in the residual layers. Also, Batch Normalization (BN) is employed to improve the training efficiency of the SMGNet. The hyper-parameters such as strides, number of kernels, and the kernel size were chosen empirically, whereas the kernel values are initialized randomly.

3.2 Skip Connected and Memory Guided Representation

The encoded representation from the last layer of the residual block is referred as $\mathbf{y_r}$, which is then fed to the memory-guided network to obtain $\hat{\mathbf{y_r}}$ as shown in Fig. 1. The memory matrix M is randomly initialized with weights of dimension NxC. N is empirically chosen to be 2000, and the dimension of C is assumed to be the same as that of $\mathbf{y_r}$. The row vector $\mathbf{m_i}$ denotes each memory item in M, where $\mathbf{m_i}$ ranges from 1 to N. The memory unit M is updated via backpropagation and gradient descent while training. During the backward pass, gradients for the memory items $\mathbf{m_i}$ which have non-zero addressing weights w_i can remain non-zero. Once an encoded representation y_r is passed into the memory-guided network, the distance of $\mathbf{y_r}$ with respect to all the memory items $\mathbf{m_i}$ is calculated as given below:

$$s(\mathbf{y_r}, \mathbf{m_i}) = \frac{\mathbf{y_r}\mathbf{m_i}^T}{\|\mathbf{y_r}\| \, \|\mathbf{m_i}\|} \tag{7}$$

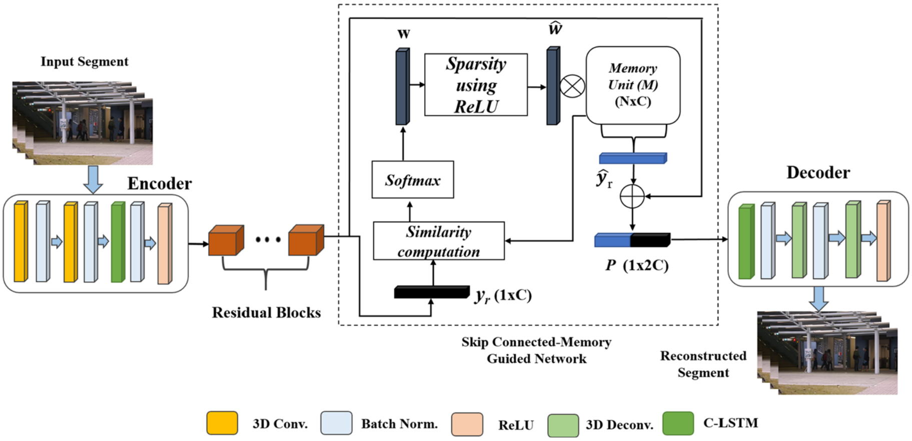


Fig. 1 Architecture of Skip Connected and Memory Guided Network (SMGNet)

Once the similarity $s(\mathbf{y_r}, \mathbf{m_i})$ is computed for the encoded representation of the test segment with every memory item, each weight w_i of the weight vector w is computed using the softmax operation as follows:

$$w_i = \frac{e^{s(\mathbf{y_r},\mathbf{m_i})}}{\sum_{j=1}^{N} e^{s(\mathbf{y_r},\mathbf{m_j})}} \tag{8}$$

Therefore, the memory-guided network redeems the memory items which are similar to $\mathbf{y_r}$, to obtain the memory-based representation $\hat{\mathbf{y}}_\mathbf{r}$ for reconstruction. After finding the weight vector w, a ReLU activation function is applied on w to obtain $\hat{w}$ for inducing sparsity. The newly updated sparse weight vector $\hat{w}$ is used to select the features from the memory matrix that represent the normality in the input frame.

The reconstructed frame will have a large margin of error when the model receives a frame that contains anomalous activity. But there is still a possibility for the calculated $\hat{\mathbf{y}}_\mathbf{r}$ to reconstruct the anomaly by combining several parts of the normality feature vectors contained in the memory matrix. This happens especially with a dense w.

One of the potential solutions is to make sure that reconstruction uses only relevant normal patterns. This can be imposed if the vector w is sparse, which is achieved based on a certain threshold chosen with respect to the size (N) of the Memory matrix M (threshold range: [1/N to 3/N]). The values in the w vector that are lesser than the threshold are made as 0, which makes the vector $\hat{w}$ sparse. One of the simpler methods of implementing this is to use a ReLU activation function to obtain $\hat{w}$.

$$\hat{w} = h(w_i; threshold) = \begin{cases} w_i, & if \quad w_i > threshold \\ 0 & \qquad otherwise \end{cases} \tag{9}$$

After the shrinkage operation, the new latent representation $\hat{\mathbf{y}}_\mathbf{r}$ is obtained using the equation,

$$\hat{\mathbf{y}}_\mathbf{r} = \sum_{i=1}^{N} \hat{w}_i \mathbf{m_i} \tag{10}$$

Since the network is forced only to store the most significant normality patterns, the reconstruction is performed only based on a small set of memory items stored in the memory.

Table 1 Architecture of the proposed R-STAE

Layer	Output-Map Dim.	Kernel	Stride	Output Channel
Image	227 × 227 × 10	–	–	–
Conv-3D 2 (tanh)	55×55×10	11×11×11	4	128
Conv-3D 3 (tanh)	26×26×10	5×5×1	2	64
C-LSTM (Conv)	26×26×10	3×3	1	64
Residual Block 1				
Conv-3D 4 (ReLU)	**26×26×10**	**3×3×1**	1	**64**
Conv-3D 5 (ReLU)	**26×26×10**	**3×3×1**	**1**	**64**
Residual Block 2				
Conv-3D 6 (ReLU)	**26×26×10**	**3×3×1**	1	**64**
Conv-3D 7 (ReLU)	**26×26×10**	**3×3×1**	**1**	**64**
Residual Block 3				
Conv-3D 8 (ReLU)	**26×26×10**	**3×3×1**	1	**64**
Conv-3D 9 (ReLU)	**26×26×10**	**3x3x1**	**1**	**64**
Memory Guided Network				
Conv.LSTM (De-Conv)	26×26×10	3×3	1	128
DeConv-3D 1(tanh)	55×55×10	5×5×1	2	128
DeConv-3D 2(tanh)	227×227×10	11×11×1	4	128

Bold indicates the configuration of residual blocks

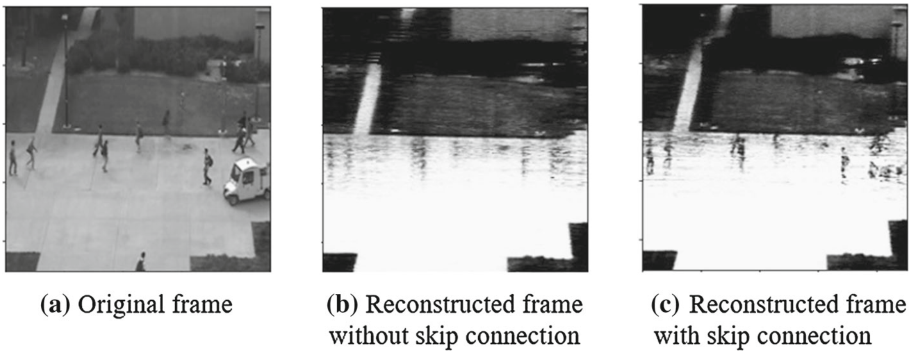
This sometimes leads to loss of information while reconstructing normal foreground objects since only a minimal set of significant normal patterns are used while reconstruction. To overcome this issue, a skip connection is also introduced in the SMGNet as shown in Fig. 1 to compensate this kind of loss of information. Using the skip connection, the encoding $\mathbf{y_r}$ obtained before the memory-guided network is concatenated to the encoding $\hat{\mathbf{y}}_\mathbf{r}$ obtained after the memory-guided network along the channel dimension to form a representation $\mathbf{P}$, and this representation is used for reconstruction by the decoder. This concatenation helps the decoder to reconstruct the incoming frames using significant normal patterns present in the memory, slightly compromising the representation capacity of the convolution layers during normality modeling.

3.3 Anomaly Detection Using SMGNet

The architecture details of the SMGNet approach are presented in Table 1. A normality model is learned using normal video segments given as input to the SMGNet network. Means Squared Error (MSE) is computed using the frame-wise difference between the reconstructed and actual frame. It becomes evident that the reconstruction error for normal frames will be higher than that of abnormal frames. The normality scores for all the frames in a segment are computed as given below, where T is the number of frames in a test segment.

$$\text{normality score} = 1 - (MSE - min(MSE_t))/max(MSE_t)), \qquad t = 1.....T \qquad (11)$$

The scores will be in the range $[0 - 1]$. Finally, a threshold value is empirically chosen and compared with the normality scores to detect the anomalous frames.



(a) Original frame

(b) Reconstructed frame without skip connection

(c) Reconstructed frame with skip connection

Fig. 2 Normal foreground objects are reconstructed well with a skip connection - Ped 2 dataset

The aim is to achieve a meaningful reconstruction of the normal video segments. During the training phase, the reconstruction error has to be minimized for normal events only through architectural stability. No pre-trained models are used in the spatio-temporal autoencoder architecture to extract the latent representations. Instead of RGB images, the SMGNet network uses grayscale images to avoid the reconstruction of unnecessary information.

The dimension of the input video segment is 227*227*1*10, where 1 denotes one channel of the gray-scale image, and 10 is the number of continuous frames forming a video segment. The effect of adding skip connections to the SMGNet network is observed in Fig. 2. The skip connection helped in achieving meaningful reconstruction of normal events without losing much spatiotemporal information. Hence, the proposed memory-guided network with skip connection is expected to improve the abnormality detection performance.

4 Experimental Studies

4.1 Datasets Used

The CUHK-Avenue dataset contains 16 training videos(15,328 frames) and 21 test videos(15,324 frames) with 47 abnormal events, which include a person walking in the wrong direction, running, throwing objects, etc. The resolution of each image is 360*640 with a frame rate of 25 frames per second (fps).

The UCSD Ped2 dataset contains 16 train videos and 12 test videos with 12 abnormal events, which include driving a vehicle, skating, riding a bike, etc. The resolution of each image is 240*360.

The Live Video (LV) dataset consists of 30 videos with unique scenarios, each containing both the train and test sequences with abnormal events such as vehicle accidents, robbery, etc. The frame rate varies from 7.5 to 30 frames per second, and its resolution varies from a minimum of 176*144 to a maximum of 1280*720.

4.2 Training

The training videos are first converted to image frames and are resized to 227 × 227. A set of 10 consecutive frames is considered as one video segment. The configuration of the

Table 2 Influence of memory guided network in the SMGNet architecture

Configuration	Avenue (AUC)	LV (AUC)
W/o memory	0.82	0.68
With memory	0.83	0.71

Table 3 Influence of skip connections in the SMGNet architecture

Configuration	Avenue (AUC)	LV (AUC)
With memory and w/o skip connection	0.83	0.71
With memory and skip connection	0.84	0.73

architecture, as shown in Table 1, is used for training. The proposed model uses Adam Optimizer with a learning rate of 0.01, and the size of the memory unit is chosen as 2000. The proposed model is implemented using Keras deep learning framework. The dataset is split into batches of size 16. All the datasets are trained for 900 epochs. The proposed model has 1,580,801 parameters. Studies were carried out with data augmentation technique reported in [14] and achieved 1% improvement over the UCSD-Ped2 dataset. But there was almost no improvement with data augmentation for Avenue and LV datasets.

Run-time The proposed SMGNet detects abnormality at 150 fps with experiments carried out on an NVIDIA QUADRO-P5000 graphics card. Anomaly detection in one frame takes only about 0.0026s, which is much faster than the previous deep learning approaches [26], [29] and [29] proving the lightweight nature of the SMGNet model.

4.3 Ablation Studies and Performance of the SMGNet

In the basic RSTAE architecture, the number of residual blocks and C-LSTM layers are empirically chosen as 3 and 2, respectively [7]. This section compares the effects of the following: (1) Effect of the memory-guided network, (2) Influence of skip connections. (3) Influence of inducing sparsity in the SMGNet architecture.

Table 2 clearly contrasts the difference in the performance of the proposed approach with and without the memory-guided network. Augmenting memory guided network improves the AUC score by 2% for the CUHK-Avenue dataset. There is an 3% improvement in the AUC score for the LV dataset, which clearly shows that the proposed model is capable of performing better with the memory guided network. As observed in Tables 5, 6, and 7, addition of memory-guided network in the R-STAE architecture improves the accuracy of from 2% to 3% for all datasets.

The effect of adding skip connections is studied for the CUHK-Avenue and LV datasets, and presented in Table 3 and in Fig. 3. It can be inferred that using only the output of the memory-guided network without a skip connection from the residual block for reconstruction resulted in the reconstruction of frames which does not completely preserve the foreground details. To ensure the reconstruction of normal foreground objects, the output of the residual layer is also appended to the output of the memory guided network. Adding skip connection shows a result improvement of 1% and 2% for CUHK-Avenue and LV datasets, respectively.

The influence of inducing sparsity is studied for UCSD-Ped2 and LV datasets, and presented in Table 4. Inducing sparsity to the memory module highly helped capturing only the

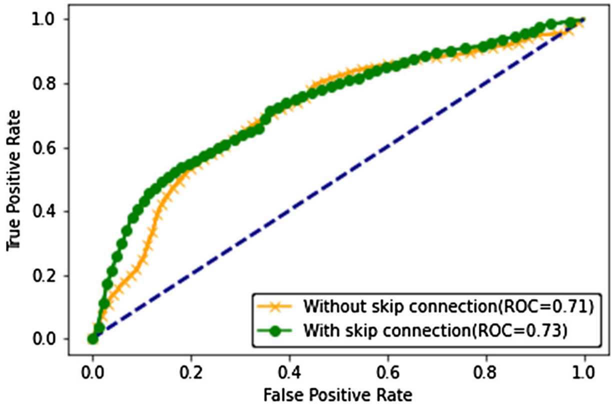


Fig. 3 ROC Curve LV - With skip connection vs Without skip connections (Scores are sampled alternatively to obtain better clarity of the curve)

Table 4 Influence of inducing sparsity in the SMGNet architecture

Configuration	UCSD-Ped2 (AUC)	LV (AUC)
With memory, skip connection and w/o sparsity	0.85	0.70
With memory, skip connection and sparsity	0.86	0.73

relevant normal patterns, which in turn improve the anomaly detection performance by 1% and 3% for UCSD-Ped2 and LV datasets respectively.

4.4 Comparison with the State-of-the-Art

Comparisons among existing VAD approaches and the SMGNet are carried out for CUHK-Avenue, LV, and Ped 2 datasets. Table 5 presents the comparison results for the CUHK-Avenue [27]. A convolutional autoencoder [14] architecture is proposed with standard HOG, HOF, and raw videos as inputs to model the spatiotemporal information with the help of reconstruction loss. Allison et al. [8] proposed a novel sliding window based discriminative learning framework for anomaly scoring. The approach is also independent of contextual assumptions of anomalies. It was able to perform quite well on the avenue dataset with an AUC of 0.78.

Another work [40] explores a convolutional winner-take-all autoencoder (CONV-WTA) with optical flow sequences as inputs to learn the normality model. The CONV-WTA approach incorporates OC-SVM instead of normality scores to detect anomalies. The ST-CaAE [24] approach detects anomalies based on a cuboid-patch-patch based technique with the optical flow as inputs to the spatiotemporal autoencoder network. Still, the approach could only achieve similar results as the SMGNet on the CUHK-Avenue dataset. The proposed Deep SMGNet approach is comparable to [26], and outperforms other state-of-the-art methods. The Frame-pred [26] approach outperforms the proposed SMGNet approach since it uses an adversarial learning framework for which the computational complexity is high compared to the proposed approach. Compared to the sRNN [29] approach, the proposed SMGNet shows a 1% increase in the AUC score.

Table 5 Performance over Avenue dataset

S. no	Method	AUC
1	Conv-Autoencoder [14]	0.70
2	Discriminative Framework [8]	0.78
3	STAE-Grayscale [45]	0.77
4	STAE-optflow [45]	0.81
5	Sparse Dictionary [27]	0.81
6	Conv-WTA+SVM [40]	0.82
7	sRNN [29]	0.82
8	ST-CaAE [24]	0.83
9	**Frame-pred** [26]	**0.85**
10	R-STAE [7]	0.82
11	MemAE [13]	0.83
12	SMGNet	0.84

Bold indicates the highest result achieved for the corresponding approach/technique

Table 6 Performance over LV dataset

S. no	Method	AUC
1	Sparse Dictionary [27]	0.11
2	H.264 [1]	0.15
3	Binary Features [21]	0.18
4	K-Means with BS [18]	0.25
5	KUGDA with BS [18]	0.26
6	Conv-Autoencoder [14]	0.34
7	Conv.LSTM-Autoencoder [28]	0.39
8	R-STAE [7]	0.68
9	**SMGNet**	**0.71**

Bold indicates the highest result achieved for the corresponding approach/technique

The LV dataset is very challenging since the context of every video is different. The SMGNet model significantly outperforms other state-of-the-art techiniques as shown in Table 6. Khan et al. [18] proposed a rejection of motion outlier approach using KUGDA (Univariate Gaussian Discriminant Analysis) for anomaly detection. Few baseline studies have been experimented by Levya et al. [22] such as [1,27]. The H.264 [1] approach was computationally less intensive, but the detection results were low since no standard techniques for feature extraction such as optical flow were not used. Since the LV dataset has videos with different scenarios, it demands a model that is capable of classifying anomalies in any general scenario. The performance of the proposed SMGNet approach is significantly better in handling varying contexts than the state-of-the-art approaches.

The UCSD-Ped2 is a small and less complex dataset when compared to the other datasets used for studies. The SMGNet approach outperformed the MPPCA+Social Force [30] approach with a 14% improvement in the AUC score. Compared to the Unmasking and R-STAE techniques, the proposed model shows a 4% and 3% increase in AUC scores, respectively. But, when compared to the other approaches in Table 7, the SMGNet is observed

Table 7 Performance over UCSD-Ped 2 dataset

S. no	Method	AUC
1	Social Force [32]	0.56
2	MPPCA+Social Force [30]	0.69
3	Unmasking [41]	0.82
4	Conv.Autoencoder [14]	0.90
4	**Abnormal GAN** [35]	**0.93**
5	R-STAE [7]	0.83
6	MemAE [13]	0.94
7	SMGNet	0.86

Bold indicates the highest result achieved for the corresponding approach/technique

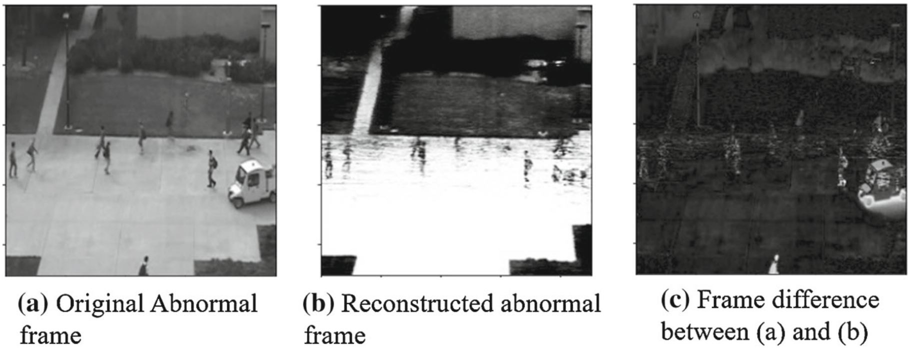

(a) Original Abnormal frame
(b) Reconstructed abnormal frame
(c) Frame difference between (a) and (b)

Fig. 4 Frame difference between the original and abnormal frame - UCSD-PED-2 dataset

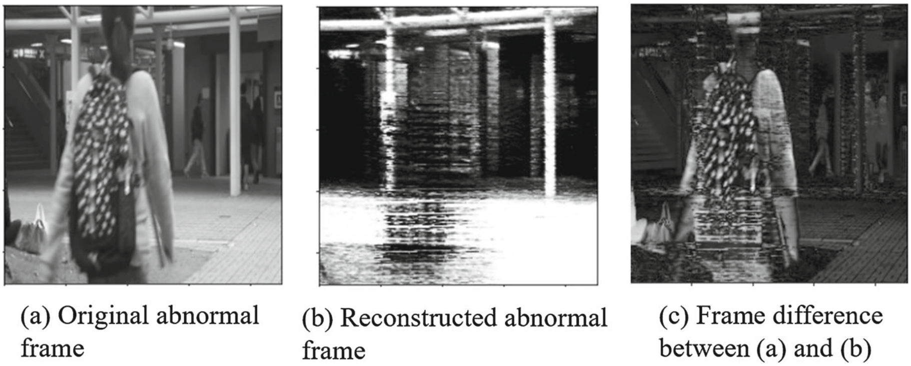
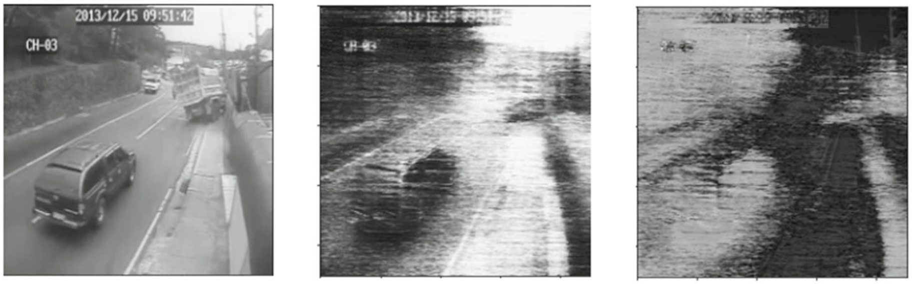

(a) Original abnormal frame
(b) Reconstructed abnormal frame
(c) Frame difference between (a) and (b)

Fig. 5 Frame difference between the original and abnormal frame- Avenue dataset

to exhibit slightly degraded performance. The AbnormalGAN [35] with Generative adversarial network as its base, is a very heavy weight model and takes more time for training and testing when compared to the proposed model. One possible justification for degraded performance of the SMGNet compared to Convolution Autoencoder [14] and MemAE [13] in UCSD-Ped2 dataset would be that the proposed approach did not augment training data in any form inspite of having smaller number of training examples in the dataset.



(a) Original abnormal frame

(b) Reconstructed abnormal frame

(c) Frame difference between (a) and (b)

Fig. 6 Frame difference between the original and abnormal frame- LV dataset

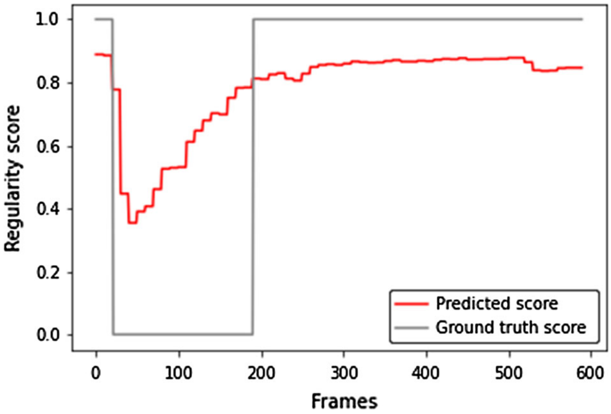


Fig. 7 Normality score-Avenue

4.5 Qualitative Analysis

The difference between the original abnormal frame and reconstructed frame from the testing sets of UCSD-Ped 2, CUHK-Avenue and LV datasets are shown in Figs. 4, 5 and, 6 respectively. In case of UCSD-Ped2 dataset, as illustrated in Fig. 4, the reconstruction of a vehicle in an abnormal frame is not clear, indicating high reconstruction error. In case of CUHK-Avenue dataset, as illustrated in Fig. 5, a person walking in wrong direction is not reconstructed properly. In the case of LV dataset, crashing of a vehicle in the anomalous frame is not reconstructed properly by the SMGNet model as shown in Fig. 6. Thus, these figures demonstrate poor reconstruction of abnormal frames and so high reconstruction error leading to effective detection of anomalous frames.

Figure 7 shows the variation of normality scores compared to the ground truth over a certain number of test frames for CUHK-Avenue dataset. The ground truth value of 1 denotes the normal frames, and the ground-truth value of 0 signifies abnormal frames. The reduction in normality score depicts higher reconstruction error of abnormal frames. The plot depicts the fact that the variation in the ground truth values and the normality score is very similar, which justifies the capability of the proposed model to discriminate between normal and abnormal frames.

5 Conclusion

In this work, we have introduced a skip connected and memory-guided network (SMGNet) for anomaly detection in videos. The addition of a memory guided network to capture and store significant normal patterns helps in the effective reconstruction of normal events so that the decoder reconstructs the abnormal events with relatively high error. Further, inducing sparsity with the help of the ReLU activation function in the memory guided network helped in achieving meaningful latent representations by using only a minimal number of memory items in the memory, which is further used for reconstruction. The addition of skip connection also helped in avoiding the loss of meaningful foreground patterns present in the input frames. Experiments on the standard benchmark datasets prove the effectiveness of the proposed approach than most of the existing state-of-the-art approaches in terms of detection performance and computational complexity.

Acknowledgements This work was supported by $No.DST/CSRI/2017/131(G)$ under Cognitive Science Research Initiative (CSRI), Department of Science and Technology, Government of India.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

References

1. Biswas S, Babu RV (2013) Real time anomaly detection in h. 264 compressed videos. In: 2013 Fourth national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG), IEEE, pp 1–4
2. Chan TH, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) Pcanet: a simple deep learning baseline for image classification? IEEE Trans Image Process 24(12):5017–5032
3. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. In: International symposium on neural networks, Springer, pp 189–196
4. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1, IEEE, pp 886–893
5. Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. In: European conference on computer vision, Springer, pp 428–441
6. D'Avino D, Cozzolino D, Poggi G, Verdoliva L (2017) Autoencoder with recurrent neural networks for video forgery detection. Electron Imaging 2017(7):92–99
7. Deepak K, Chandrakala S, Mohan CK (2021) Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Sig Image Video Process 15(1):215–222
8. Del Giorno A, Bagnell JA, Hebert M (2016) A discriminative framework for anomaly detection in large videos. In: European conference on computer vision, Springer, pp 334–349
9. Dong F, Zhang Y, Nie X (2020) Dual discriminator generative adversarial network for video anomaly detection. IEEE. Access
10. Feng Y, Yuan Y, Lu X (2017) Learning deep event models for crowd anomaly detection. Neurocomputing 219:548–556
11. Fritzke B (1995) A growing neural gas network learns topologies. In: Advances in neural information processing systems, pp 625–632
12. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
13. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, Hengel Avd (2019) Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE international conference on computer vision, pp 1705–1714
14. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 733–742

15. Hinami R, Mei T, Satoh S (2017) Joint detection and recounting of abnormal events by learning deep generic knowledge. In: Proceedings of the IEEE international conference on computer vision, pp 3619–3627
16. Ionescu RT, Khan FS, Georgescu MI, Shao L (2019) Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7842–7851
17. Ionescu RT, Smeureanu S, Popescu M, Alexe B (2019) Detecting abnormal events in video using narrowed normality clusters. In: 2019 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1951–1960
18. Khan MUK, Park HS, Kyung CM (2018) Rejecting motion outliers for efficient crowd anomaly detection. IEEE Trans Inf Forensics Secur 14(2):541–556
19. Kratz L, Nishino K (2009) Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 1446–1453
20. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA et al (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis 123(1):32–73
21. Leyva R, Sanchez V, Li CT (2017) Abnormal event detection in videos using binary features. In: 2017 40th international conference on telecommunications and signal processing (TSP), IEEE, pp 621–625
22. Leyva R, Sanchez V, Li CT (2017) The lv dataset: A realistic surveillance video dataset for abnormal event detection. In: 2017 5th international workshop on biometrics and forensics (IWBF), IEEE, pp 1–6
23. Li N, Chang F (2019) Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder. Neurocomputing 369:92–105
24. Li N, Chang F, Liu C (2020) Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Trans Multimed 23:203–215
25. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, pp 740–755
26. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6536–6545
27. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision, pp 2720–2727
28. Luo W, Liu W, Gao S (2017) Remembering history with convolutional lstm for anomaly detection. In: 2017 IEEE international conference on multimedia and Expo (ICME), IEEE, pp 439–444
29. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked rnn framework. In: Proceedings of the IEEE international conference on computer vision, pp 341–349
30. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, pp 1975–1981
31. Medel JR (2016) Anomaly detection using predictive convolutional long short-term memory units. Thesis. Rochester Institute of Technology
32. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 935–942
33. Van den Oord A, Schrauwen B (2014) Factoring variations in natural images with deep gaussian mixture models. In: Advances in neural information processing systems, pp 3518–3526
34. Ramachandra B, Jones M, Vatsavai R (2020) Learning a distance function with a siamese network to localize anomalies in videos. In: The IEEE winter conference on applications of computer vision, pp 2598–2607
35. Ravanbakhsh M, Nabi M, Sangineto E, Marcenaro L, Regazzoni C, Sebe N (2017) Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE international conference on image processing (ICIP), IEEE, pp 1577–1581
36. Shi Y, Tian Y, Wang Y, Huang T (2017) Sequential deep trajectory descriptor for action recognition with three-stream cnn. IEEE Trans Multimed 19(7):1510–1520
37. Srivastava N, Mansimov E, Salakhudinov R (2015) Unsupervised learning of video representations using lstms. In: International conference on machine learning, pp 843–852
38. Sun Q, Liu H, Harada T (2017) Online growing neural gas for anomaly detection in changing surveillance scenes. Pattern Recogn 64:187–201
39. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
40. Tran HT, Hogg D (2017) Anomaly detection using a convolutional winner-take-all autoencoder. In: Proceedings of the British machine vision conference 2017. British Machine Vision Association

41. Tudor Ionescu R, Smeureanu S, Alexe B, Popescu M (2017) Unmasking the abnormal events in video. In: Proceedings of the ieee international conference on computer vision, pp 2895–2903
42. Xingjian S, Chen Z, Wang H, Yeung DY, Wong WK, Woo Wc (2015) Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, pp 802–810
43. Xu D, Ricci E, Yan Y, Song J, Sebe N (2015) Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553
44. Zhao Y, Deng B, Shen C, Liu Y, Lu H, Hua XS (2017) Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM international conference on Multimedia, pp 1933–1941
45. Zhao Y, Deng B, Shen C, Liu Y, Lu H, Hua XS (2017) Spatio-temporal autoencoder for video anomaly detection. In: ACM Multimedia