Signal, Image and Video Processing (2022) 16:2001–2007
https://doi.org/10.1007/s11760-022-02161-y

ORIGINAL PAPER

Object-centric and memory-guided network-based normality modeling for video anomaly detection

S. Chandrakala[1] · P. Shalmiya[1] · V. Srinivas[1] · K. Deepak[1]

Received: 13 May 2021 / Revised: 13 January 2022 / Accepted: 21 January 2022 / Published online: 19 February 2022

Abstract
Anomaly detection in surveillance videos is a challenging and demanding task. Autoencoders trained on segments of normal events are expected to give high reconstruction error for abnormal events than that for normal events. However, the assumption of autoencoders giving high reconstruction error is not always true in practice. Since the autoencoder sometimes offers better generalization, it also reconstructs abnormal events well, leading to slightly degraded performance for anomaly detection. Another issue is that the performance of real-time anomalous activity detection in surveillance videos still needs improvement. To address these issues, we propose an Object-centric and Memory-guided residual spatiotemporal autoencoder (OM-RSTAE) to detect video anomalies. The proposed technique achieved improved results over benchmark datasets, namely UCSD-Ped2, Avenue, ShanghaiTech and UCF-Crime datasets.

Keywords Video anomaly detection · Normality modeling · Memory guided network · Spatiotemporal autoencoders · Residual blocks · Anomalous objects

1 Introduction

Real-time video anomaly detection (VAD) systems are of great demand to ensure private and public safety. There exist various challenges while building a VAD system. Insufficient training data for anomalous activities create an imbalance between normal and anomalous samples. Data points lie in a higher dimension and formulation of anomalies may differ based on scenarios, e.g., running in the middle of the road might be considered anomalous while running in a park is not. VAD can be posed as an outlier detection problem in which a normality model is built using data of normal activities. While testing, any deviations from characteristics learned by the normality model are recognized as anomalies. Initially, reconstruction error-based autoencoders such as Hasan et al. [5] solely depend on a 2D convolutional autoencoder in which the convolution and pooling operations are performed only in the spatial dimensions, and so it fails to capture the temporal characteristics. To overcome this issue, 3D-convolution layers and convolutional LSTM (C-LSTM) layers are augmented to the autoencoder for modeling motion information [19].

Recently, we proposed a Residual Spatio-temporal Autoencoder (R-STAE) [1]-based normality modeling approach to learn the spatio-temporal information present in the video segments. An important issue of normality model-based approaches is that autoencoders may provide better generalization so that few anomalous activities might also be reconstructed well. Another issue is that the performance of real-time anomalous activity detection in surveillance videos still needs improvement. We propose an Object-centric and Memory-guided residual spatiotemporal autoencoder (OM-RSTAE)-based normality modeling approach to detect video anomalies as an extension to our R-STAE-based approach. We explore the memory-module used in MemAE [4] approach to capture the significant normality patterns present in the training data. MemAE approach uses Convolutional Autoencoder (CAE) along with memory-module

✉ S. Chandrakala
chandrakala@cse.sastra.edu

P. Shalmiya
shalmiya@cse.sastra.ac.in

V. Srinivas
srinivasvasudevan2000@gmail.com

K. Deepak
deepak@sastra.ac.in

[1] Intelligent Systems Group, School of Computing, SASTRA Deemed to be University, Thanjavur, Tamil Nadu 613401, India

to model the normality patterns, whereas in this proposed work, we augment the memory-module in R-STAE architecture to detect anomalies in surveillance videos. In addition, the proposed architecture initially detects anomalous objects in the video using a pre-trained object detection model as the first level. The anomalies which are not detected in the first level are further processed using Memory-guided R-STAE architecture to identify the temporal anomalies. The overall result of the proposed approach is the weighted average of first-level detection and second level detection.

2 Related work

So far in the literature, the techniques proposed for VAD fall under two categories: (1) modeling events using hand-crafted feature-based techniques which make use of features such as histogram of gradients, 3D-gradients, histogram of optical flow, trajectories [15], 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. A non-deep state-of-the-art approach for such unsupervised modeling involves a combination of sparse coding, and bag-of-words [9]. However, bag-of-words do not preserve the spatio-temporal structure of the words and require prior information about the number of words. Additionally, optimization involved in sparse coding for both training and testing is computationally expensive, especially with large data such as videos.

3D-Convolution architectures used to design the 3D autoencoders to obtain high-level features are invariant to intra-class spatiotemporal changes [20]. This approach uses stacked frames as an input to the 3D-filters as done in Fully connected AE [5] approach. The feature maps obtained out of 3-D filters are used to model the spatiotemporal changes. A prediction stream is also used to better handle the issue of poorly reconstructed normal events by the autoencoder stream. Local temporal coherence was taken into consideration while designing the prediction loss. A semi-supervised learning approach for VAD using dual discriminator-based GAN architecture is proposed in [3]. 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 or fake.

3 Object-centric and memory-guided-RSTAE

As shown in Fig. 1, Object-centric and Memory-guided Residual Spatiotemporal Autoencoder (OM-RSTAE) is proposed to detect anomalous objects in the first level using a pre-trained object detection model. The anomalies which are not detected in the first level are further processed using Memory-guided R-STAE architecture to identify the temporal anomalies. In real-time, pre-trained object detection models can be used to detect anomalous objects in the sequence of frames in scenarios such as pedestrian walkways and campus scenarios. In addition to the anomalous object detection model, augmenting the memory module in the R-STAE architecture [1] helps in memorizing the significant normality patterns present in the training data of normal activities.

3.1 Anomalous object detection

Detecting anomalous objects as the first level simplifies the VAD system with improved efficacy. A pretrained object detection model trained on the COCO-17 dataset, taken from Tensorflow object detection model zoo [13] is used in the anomalous object detection module. The EfficientDet D7 is chosen because of its high mean average precision (mAP) score of 51.2 compared to all the other pretrained models present in the tensorflow 2 model zoo. EfficientDet D7 uses BiFPN (Bi-directional Feature Pyramid Network), a bidirectional feature network that takes input from multiple layers of the EfficientNet backbone. The BiFPN uses the multi-level feature fusion technique.

The efficiency is further increased by using a fast normalized fusion technique that takes into consideration the effects of input features at different resolutions having unequal contribution to the output features. Therefore, appropriate additional weights are assigned that allow the network to learn the importance of different input features. As a result of which there is a significant 4% increase in accuracy and 50% reduction in computational cost compared to the previously used feature fusion technique in the NAS-FPN [13] architecture. The fused features are then used by the class/box network to predict the location and class of each object. A set of image frames with channels, $n*h*w*c$, are given as input to the pretrained object detection model as shown in Fig. 1. The model identifies all the objects in the frame and the identified objects are stored in a list. The objects in the list are compared against a predefined list of anomalous objects to identify whether those set of images are anomalous. Further, the input frames are fed into the Memory-guided R-STAE architecture. The anomalous object detection model helped in finding about 87.62 and 48.03% anomalous frames present in the testing videos of UCSD-Ped-2 and Shanghaitech Dataset, respectively.

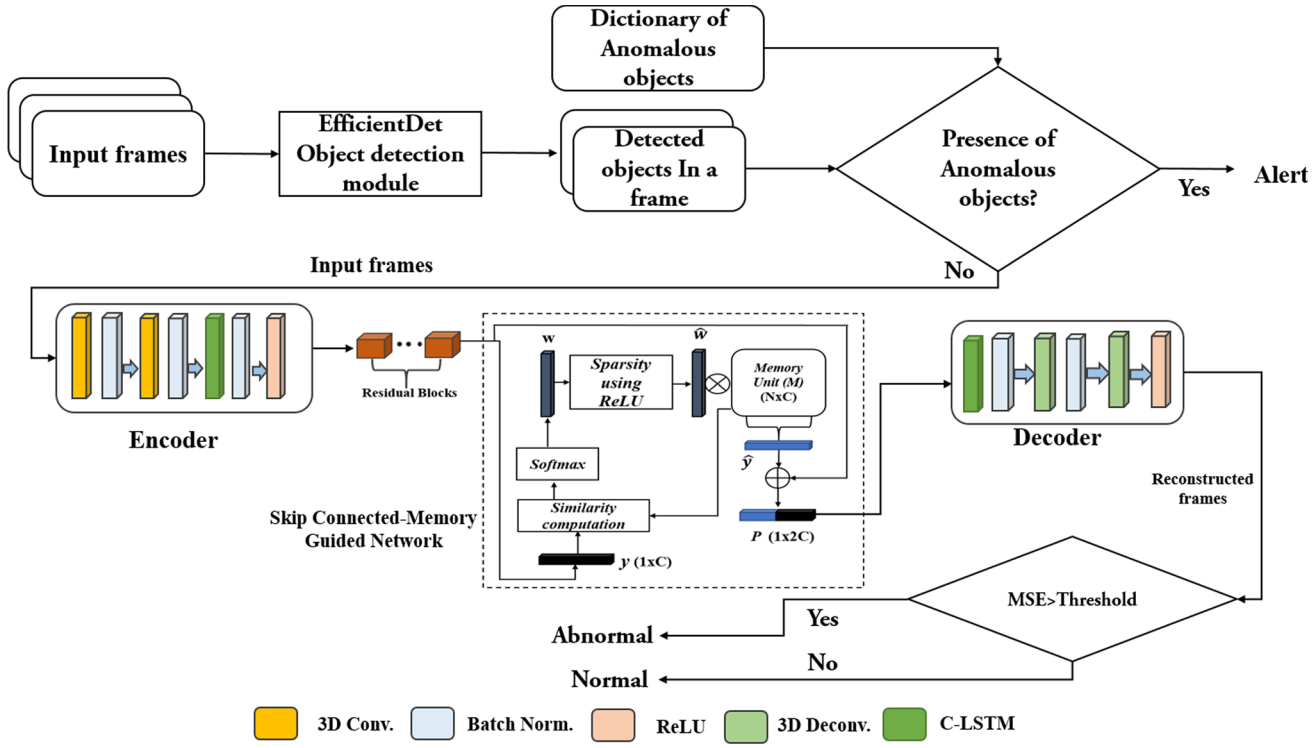


Fig. 1 Object-centric and Memory-guided Residual spatiotemporal Autoencoder (OM-RSTAE)-based approach

3.2 Skip connected and memory-guided network

To detect the unnoticed anomalies in the first level, the input frames, $n * h * w * c$, are again passed to the skip connected and Memory guided R-STAE as shown in Fig. 1. The encoder consists of two 3-D convolution layers and then two Convolutional-LSTM (C-LSTM) layers. 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. The configuration of residual blocks used in the proposed OM-RSTAE architecture 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 bestows for gradient flow while backpropagation, thus helping in taking control over the vanishing gradients. The equation of a residual block with input z is given by,

$$\mathbf{y} = F(z) + z \tag{1}$$

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

The encoded representation from the last layer of the residual block is referred as $\mathbf{y}$, which is then fed to the memory-guided network to obtain $\hat{\mathbf{y}}$ 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 $\mathbf{y}$. 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 which has nonzero weights can remain nonzero. Once an encoded representation y is passed into the memory-guided network, the distance of $\mathbf{y}$ with respect to all the memory items $\mathbf{m_i}$ is calculated as given below:

$$s(\mathbf{y}, \mathbf{m_i}) = \frac{\mathbf{y}\mathbf{m_i}^T}{\|\mathbf{y}\| \, \|\mathbf{m_i}\|} \tag{2}$$

Once the similarity $s(\mathbf{y}, \mathbf{m_i})$ 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.

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**	**3 × 3 × 1**	**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

conv-3D bold to denote that this comes under Residual block 1, 2 and 3

$$w_i = \frac{e^{s(\mathbf{y},\mathbf{m_i})}}{\sum_{j=1}^{N} e^{s(\mathbf{y},\mathbf{m_j})}} \tag{3}$$

Therefore, the memory-guided network retrieves the memory items which are similar to $\mathbf{y}$, to obtain the memory-based representation $\hat{\mathbf{y}}$ 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 memory items that represent the normality patterns.

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 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 by a sparse w, which is achieved based on a certain threshold chosen with respect to the size (N) of the Memory matrix M (threshold range: [1–3/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 (h) to obtain $\hat{w}$.

$$\hat{w} = h(w_i; threshold) = \begin{cases} w_i, & \text{if } w_i > \text{ threshold} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

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

$$\hat{\mathbf{y}} = \sum_{i=1}^{N} \hat{w}_i \mathbf{m_i} \tag{5}$$

Since the network is forced to store only the most significant normality patterns, the reconstruction is performed based on a small set of memory items stored in the memory. 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, an additional skip connection is also introduced from the output residual blocks to the output of the memory module in the OM-RSTAE as shown in Fig. 1 to compensate for this kind of loss of information. Using the skip connection, the encoding $\mathbf{y}$ is concatenated to the encoding $\hat{\mathbf{y}}$ 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.

Apart from the anomalous object detection model as the first level, the architecture details of the skip connected and memory-guided network in the OM-RSTAE approach are presented in Table 1. The normality model is learned using normal video segments given as input to the OM-RSTAE model. The normality score is computed based on the MSE

Table 2 Run-time analysis

Method	Time taken (One frame) (s)
Frame-pred [8]	0.12
Sparse-coding [10]	0.02
OM-RSTAE	0.0026

values obtained by computing the frame-wise difference between the reconstructed and actual frame. The normality 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 segments at the second level. The overall result of the proposed approach is the weighted average of first-level detection and second level detection.

4 Experimental studies

4.1 Datasets

We conducted experiments on the following datasets: CUHK Avenue [9], Shanghaitech [8], UCSD-Ped 2 [11] and UCF-Crime [16]. 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 UCSD Ped2 dataset contains 16 training videos and 12 test videos with 12 abnormal events, which include driving a vehicle, skating, riding a bike, etc. The ShanghaiTech Campus dataset has 13 scenes with complex light conditions and camera angles. It contains 130 abnormal events and over 270, 000 training frames. UCF-Crime dataset consists of about 13 activities describing the real-world anomalies. The dataset has a total of 800 normal video sequences for training and 290 sequences for testing. The Area Under the Curve (AUC) scores are used as an effective metric to evaluate the performance since the ratio between normal and abnormal events in test video is not similar.

4.2 Training and Ablation studies

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 dataset is split into batches of size 16 for training. *Run-time:* the proposed model has 1,580,801 parameters. The proposed OM-RSTAE detects abnormality at 150 fps with experiments carried out on an NVIDIA QUADRO-P5000 graphics card. As shown in Table 2, anomaly detection in one frame takes only about 0.0026 which is much faster than the deep learning approaches [8,10].

Table 3 Influence of memory-guided network in the OM-RSTAE architecture

Configuration	Avenue (AUC)	SHANGHAITECH (AUC)
W/o skip connected memory module	0.82	0.68
With skip connected memory module	0.83	0.71

The number of residual blocks and C-LSTM layers in the base R-STAE architecture [1] are empirically chosen as 3 and 2, respectively. Table 3 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 1% for the CUHK-Avenue dataset. There is an 3% improvement in the AUC score for the SHANGHAITECH dataset, which clearly shows that the proposed model is capable of performing better with the memory-guided network.

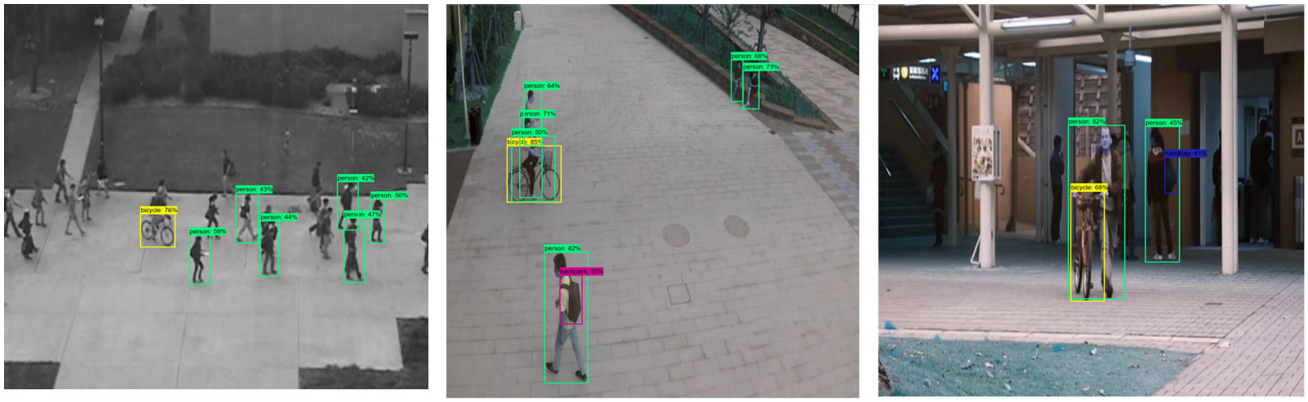
4.3 Qualitative analysis

The performance of the EfficientDet D7 object detection model [13] is illustrated in Fig. 2. The detection of anomalous object, which is a bicycle in the pedestrian pathway from 3 datasets is presented. As seen in Fig. 2b, even though the anomalous bicycle is occluded by pedestrians, the pre-trained object detection model in the first level is capable of detecting the bi-cycle object. The anomalous object detection model helped in finding about 87.62 and 48.03% of anomalous frames present in the testing videos of UCSD-Ped-2 and SHANGHAITECH dataset, respectively. The overall result of the proposed approach is the weighted average of first-level detection (pre-trained object detection model) and second level detection (Skip-connected and memory-guided module). The weighted average results are reported as the overall results of proposed approach.

4.4 Performance analysis

Table 4 presents the comparison results for the CUHK-Avenue dataset [9]. Allison et al. [2] proposed a novel sliding window-based discriminative learning framework for anomaly scoring. The approach is also independent over contextual assumptions of anomalies. It was able to perform quite well on the avenue dataset with an AUC of 0.78. A convolutional autoencoder [5] architecture is proposed with standard HOG, HOF, and raw videos as inputs to model the spatiotemporal information with the help of reconstruction loss.

Another work [17] explores a convolutional winner -take-all autoencoder (CONV-WTA) with optical flow sequences



(a) (b) (c)

Fig. 2 Qualitative analysis-Bi-cycle Object detected in **a** UCSD-Ped2, **b** Shanghaitech , and **c** Avenue dataset

Table 4 Performance over Avenue dataset

S. no	Method	AUC
1	Discriminative framework [2]	0.78
2	Conv-autoencoder[5]	0.70
3	STAE-grayscale [20]	0.77
4	STAE-optflow [20]	0.81
5	Sparse dictionary [9]	0.81
6	Conv-WTA+SVM [17]	0.82
7	sRNN [10]	0.82
8	ST-CaAE [7]	0.83
9	Frame-pred [8]	**0.85**
10	R-STAE [1]	0.82
11	MemAE [4]	0.83
12	OM-RSTAE	0.84

Bold denotes max result

Table 5 Performance over SHANGHAITECH dataset

S. no	Method	AUC
1	Conv-Autoencoder[5]	0.61
2	sRNN [4]	0.68
3	MemAE [1]	0.71
4	**Frame pred** [8]	**0.73**
5	R-STAE	0.66
6	OM-RSTAE	0.71

Bold denotes max result

Table 6 Performance over UCSD-Ped 2 dataset

S. no	Method	AUC
1	Social force [12]	0.56
2	MPPCA+social force [11]	0.69
3	Unmasking[18]	0.82
4	Conv.autoencoder [5]	0.90
5	Narrowed normality clusters [6]	0.89
6	Abnormal GAN [14]	0.93
7	R-STAE [1]	0.83
8	MemAE [4]	0.94
9	OM-RSTAE	**0.94**

Bold denotes max result

Table 7 Performance over UCF-crime dataset

S. no	Method	AUC
1	Deep autoencoder [5]	0.56
2	Lu et al [9]	0.57
3	R-STAE [1]	0.64
4	**OM-RSTAE**	**0.65**

Bold denotes max result

as inputs to learn the normality model. The CONV-WTA approach incorporates OC-SVM instead of normality scores to detect anomalies. The ST-CaAE [7] approach detects anomalies based on a cuboid-patch-based cascading technique with the optical flow as inputs to the spatiotemporal autoencoder network. Still, the approach could only achieve similar results as the OM-RSTAE on the CUHK-Avenue dataset. Compared to the sRNN [10] approach, the proposed OM-RSTAE shows a 2% increase in the AUC score. The Frame-pred [8] approach performs comparable to the proposed approach since it uses an adversarial learning framework for which the computational complexity is high compared to the proposed approach.

In case of SHANGHAITECH dataset, the Frame-pred [8] approach has achieved 2% improvement over proposed approach as shown in Table 5 .The Frame-pred [8] approach uses additional modules for estimating optical flow, which requires more network parameters and groundtruth flow fields. Moreover, Frame-Pred leverages an adversarial learning framework, taking lots of effort to train the network. On the contrary, our model uses a simple skip-connected and Memory augmented R-STAE for extracting features and detecting the anomalies.

In case of UCSD-Ped2, OM-RSTAE approach outperformed the MPPCA+Social Force [11] approach with a 25% improvement in the AUC score as shown in Table 6. Compared to the Unmasking [18] and R-STAE [1] techniques, the proposed model shows a 12 and 11% increase in AUC scores, respectively. The AbnormalGAN [14] approach achieved 0.93% AUC with generative adversarial network as its base, which is a heavyweight model and takes more time for training and testing when compared to the proposed model.

The UCF-Crime dataset [16] is challenging since the training video sequences and the corresponding testing video sequences are from different scenes. UCF-Crime dataset has data for both normal and abnormal events for training. We experimented UCF-Crime dataset in an outlier detection fashion (i.e.,) used only normal events for training, and compared with approaches that follow the same approach and is presented in Table 7. Still, the proposed was able to significantly better than the existing frame-pred [5] and sparse coding[9] approaches as shown in Table 7.

5 Conclusion

We have introduced an Object-centric and Memory-Guided Residual Spatiotemporal Autoencoder (OM-RSTAE) for anomaly detection in videos. The anomalous object detection model as the first-level, helped in identifying the anomalous objects beforehand. The addition of a skip-connected and memory-guided network to capture and store significant normal patterns helped in the effective reconstruction of normal events so that the decoder reconstructs the abnormal events with relatively high error. Further, inducing sparsity in the memory-guided network helped in achieving meaningful latent representations using only a minimal number of patterns in the memory unit which is further used for reconstruction.

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 Authors declare that they have no conflict of interest.

References

1. Deepak, K., Chandrakala, S., Mohan, C.K.: Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image Video Process **15**, 215–222 (2021)
2. Del Giorno, A., Bagnell, J. A., Hebert, M.: A discriminative framework for anomaly detection in large videos. In: European Conference on Computer Vision. Springer, pp 334–349 (2016)
3. Dong, F., Zhang, Y., Nie, X.: Dual discriminator generative adversarial network for video anomaly detection. IEEE Access **8**, 88170–88176 (2020)
4. Gong, D., Liu, L., Le, V., et al.: 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 (2019)
5. Hasan, M., Choi, J., Neumann, J., et al.: Learning temporal regularity in video sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 733–742 (2016)
6. Ionescu, R.T., Smeureanu, S., Popescu, M., et al.: Detecting abnormal events in video using narrowed normality clusters. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp 1951–1960 (2019)
7. Li, N., Chang, F., Liu, C.: Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Trans Multimedia **23**, 203–215 (2021)
8. Liu, W., Luo, W., Lian, D., et al.: Future frame prediction for anomaly detection—a new baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6536–6545 (2018)
9. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2720–2727 (2013)
10. Luo, W., Liu, W., Gao, S.: 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 (2017)
11. Mahadevan, V., Li, W., Bhalodia, V., et al.: Anomaly detection in crowded scenes. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, pp 1975–1981 (2010)
12. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp 935–942 (2009)
13. Pkulzc, R.V., Neal, W.: Tensorflow detection model zoo. GitHub [Online] (2019). https://github.com/tensor-flow/models/blob/master/research/object.detection/g3doc/detection.model.zoo.md
14. Ravanbakhsh, M., Nabi, M., Sangineto, E., et al.: Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE International Conference on Image Processing (ICIP). IEEE, pp 1577–1581 (2017)
15. Shi, Y., Tian, Y., Wang, Y., et al.: Sequential deep trajectory descriptor for action recognition with three-stream CNN. IEEE Trans. Multimedia **19**(7), 1510–1520 (2017)
16. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 6479–6488 (2018)
17. Tran, H.T., Hogg, D.: Anomaly detection using a convolutional winner-take-all autoencoder. In: Proceedings of the British Machine Vision Conference 2017. British Machine Vision Association (2017)
18. Tudor Ionescu, R., Smeureanu, S., Alexe, B., et al.: Unmasking the abnormal events in video. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2895–2903 (2017)
19. Zhao, Y., Deng, B., Shen, C., et al.: Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM International Conference on Multimedia, pp 1933–1941 (2017)
20. Zhao, Y., Deng, B., Shen, C., et al.: Spatio-temporal autoencoder for video anomaly detection. In: ACM Multimedia (2017)