Parallelizing Video Anomaly Detection Using Reconstruction and Future Frame Prediction

Vibhav Vasudevan, Srinivas Ramakrishnan, Utkarsh Seth, Shreya M B, and Shylaja SS

PES University
Center for Data Science and Applied Machine Learning
RR Campus, Bengaluru, Karnataka, India

Abstract. Video anomaly detection (VAD) is a demanding task because the very definition of anomalies in videos is inherently inconclusive and also due to the high manpower required to supervise lengthy videos. This research paper introduces a novel method for anomaly detection in videos. It utilizes the concurrent output of two deep learning models: the Convolutional Autoencoder (ConvAE) for anomaly detection based on reconstruction errors and the Convolutional Long Short-Term Memory (ConvLSTM) for future frame prediction. The ConvAE detects anomalies by capitalizing on its excellent spatial learning capabilities and the ConvLSTM model is helpful owing to its powerful temporal modeling abilities. By running these two models in parallel and normalizing the results obtained from both, we found that our combined model (CAEL-STM) gave satisfactory results (AUROC) for two of the most prevalent datasets in this field of VAD, namely CUHK Avenue (77.44%) and Ped2 (87.31%), showcasing its promising performance.

Keywords: reconstruction error, temporal modeling, spatial learning.

1 Introduction

Video Anomaly Detection (VAD) in real-time surveillance videos is a critical area of research with widespread applications in enhancing security and safety. The ability to automatically identify abnormal events and behaviors in surveillance footage plays a pivotal role in ensuring timely responses to potential threats and improving overall surveillance effectiveness.

Traditional methods of VAD often rely on handcrafted features and predefined rules, making them less adaptive and unable to cope with complex and dynamic real-world scenarios. In recent years, deep learning approaches have shown great promise in addressing the challenges of video anomaly detection. Among these approaches, Convolutional Autoencoder (ConvAE) and Convolutional Long Short-Term Memory (ConvLSTM) has gained significant attention due to their remarkable capabilities in learning spatial features and modeling temporal dependencies, respectively.

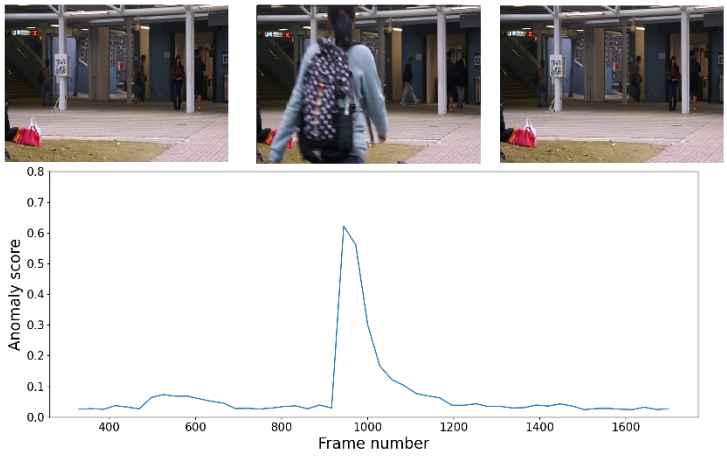


Fig. 1: This image shows the performance of our combined model on video 1 of the CUHK Avenue [12] dataset with anomaly score on the y-axis and frame number on the x-axis. As can be seen, the anomaly score rises high when the girl blocks the camera and falls back down when she leaves the sight of the mounted camera.

The research paper's problem statement centers on the creation of a real-time video anomaly detection system, harnessing the synergistic capabilities of Convolutional Autoencoder for reconstruction and Convolutional Long Short Term Memory (LSTM) for frame prediction. A key challenge is achieving precise identification of abnormal events in videos during real-time frame processing for timely alerts and responses. Notably, our novelty lies in introducing parallelization to enhance the collaboration between LSTM and Autoencoder, avoiding a sequential approach thus ensuring efficient utilization of both methods. The shortcomings of existing VAD approaches often stem from the inability to effectively model the intricate spatial and temporal patterns inherent in surveillance videos. Handcrafted features may not capture the full complexity of anomalies, leading to limited detection performance. Additionally, some deep learning-based methods focus solely on spatial or temporal aspects, neglecting the importance of considering both in an integrated manner.

This paper introduces a hybrid approach that integrates the strengths of Convolutional Autoencoder and Convolutional Long Short-Term Memory. Our approach's novelty lies in parallelizing results between LSTM and Autoencoder through result normalization. The Convolutional Autoencoder excels at learning spatial features and reconstructing normal frames accurately. By comparing the reconstruction errors, our model can effectively detect anomalous frames that deviate from normal patterns.

In contrast, the Convolutional Long Short-Term Memory specializes in modeling temporal dependencies and predicting future frames based on historical context. Leveraging its ability to distinguish normal frames from anomalous

ones over time, our approach reinforces the anomaly detection capability of the Convolutional Autoencoder, leading to improved detection accuracy.

CAE and CLSTM are both effective deep learning architectures in anomaly detection tasks. The Convolutional Autoencoder learns hierarchical spatial features, facilitating the reconstruction of normal frames and anomaly identification through error analysis. Conversely, Convolutional Long Short-Term Memory, being recurrent, captures temporal dependencies and predicts future frames, offering context for anomaly detection. Results from both models are normalized, and conclusions are drawn.

In the research paper, we conduct a comprehensive evaluation of our proposed video anomaly detection system using the Area Under the Receiver Operating Characteristic Curve (AUROC). A higher AUROC score indicates better discrimination between normal and anomalous frames. It serves as a quantitative measure of the model's overall detection performance. We conduct experiments on two widely used datasets, namely CUHK Avenue Dataset and Ped2 Dataset to demonstrate the efficacy of this hybrid system in achieving superior detection performance, laying the foundation for more adaptive surveillance systems.

2 Literature Review

There has been a large development in video anomaly detection (VAD) methods, especially with the recent development of deep neural networks. Traditionally, methods used comprised object detection [2], and feature extraction where anomalies were scored frame by frame. A comprehensive review of current VAD techniques by Abbas. et al [1] showed that currently, reconstruction and prediction-based techniques are both very prevalent and also very successful. Nassif et. al [16] found that machine learning techniques are at the forefront of VAD, with them achieving State-of-the-art results on the most popular datasets in this field. Most current methods use reconstruction or prediction-based methods attempting to capture both spatial and temporal features in videos and generate an anomaly score to classify anomalies.

Reconstruction based methods. In typical reconstruction-based methods, an autoencoder is used which is trained on normal videos only to help it to learn normal patterns. The assumption made is that this autoencoder trained solely on normal frames will not be able to reconstruct anomalous frames as well, and then a threshold can be said to classify frames into anomalous and non-anomalous. [4], [11] have also used incorporated a memory aspect into autoencoders and have been shown to achieve state-of-the-art results. However, this autoencoder approach has resulted in a few issues as at times the autoencoders reconstruct anomalies also very well. Some related works include [4], [11], [17], [22].

Frame prediction-based methods. Frame prediction-based approaches assume that anomalous events cannot be predicted appropriately since they deviate from expected behavior. Optical flow has shown some promise in this regard as it captures the temporal coherence very well and it helps estimate the mo-

tion of objects between consecutive frames. Since anomalous events are assumed to have unexpected motion patterns, they can be detected using the deviated optical flow. Some works include [10], [11], [18]. With the introduction of ConvLSTM [19], the ability to capture both temporal and spatial coherence was significantly enhanced and very popular in the field of VAD. LSTMs used in combination with another paradigm [13], [14] have achieved state-of-are results on certain benchmark datasets. Notable works include [13], [14], [15].

Other methods. Georgescu et. al [3] proposed a novel object-level approach using multi-task learning and semi-supervised learning. They utilized a pre-trained detector and trained a 3-D convolutional neural network to model temporal dependencies. Ingle et al. [6] proposed an object detection method based on MSD-CNN (Multiclass Subclass CNN) to locate and classify different types of guns and knives (subclass) in live video feed. Liu et al. [9] proposed a Diversity-Measurable Anomaly Detection (DMAD) framework to enhance reconstruction diversity while avoiding the undesired generalization of anomalies. They use deformation fields to model the motion and behavior of different anomaly types. Reiss. et al. [18] classifies anomalies into three different categories, velocity, deep, and pose and they compute anomalies with feature extraction and density estimation. Sultani et al.[20] proposed a deep multiple instance learning (MIL) framework, treating normal and anomalous videos as bags and short segments as instances to avoid annotating the anomalous segments or clips in training videos, which is very time-consuming. Tran et al. [21] suggested a spatiotemporal feature learning approach using pre-trained deep 3-dimensional convolutional networks. The extracted Conv-3D features are given as input to a multi-class linear SVM for training models for action recognition and anomaly recognition.

We propose a combination of reconstruction and frame prediction methods to leverage the powerful capabilities of both approaches.

3 Methodology

Our proposed solution to the video anomaly detection problem attempts to identify anomalous frames in a video by using frame reconstruction and future frame prediction. This is done by making use of two deep neural architectures in parallel: an Autoencoder model for frame reconstruction and a Convolutional Long Short Term Memory (ConvLSTM) model for future frame prediction. The autoencoder model reconstructs a single frame by using convolution layers in an encoder and transposed convolution layers in a decoder. The reconstructed frame is compared with the actual frame to obtain some loss, which is one of the parameters used to decide whether the frame is anomalous or not. The ConvLSTM model predicts a future frame. This predicted future frame is compared with the actual future frame to obtain a loss value which is the other parameter. These two parameters are coupled together to finally classify the frame as anomalous or non-anomalous. Our proposed solution attempts to tackle general anomaly detection: we aim to identify any behavior that is different from the norm and classify that as an anomaly. To achieve this, both the Autoencoder and ConvL-

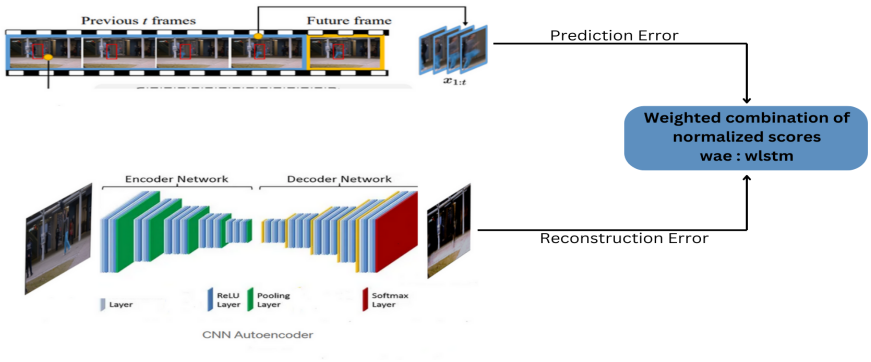


Fig. 2: This figure shows a complete overview of our combined Conv-AE and Conv-LSTM models. The reconstruction error generated by the Conv-AE model is combined with the frame prediction error generated by the Conv-LSTM model using z-score normalization and the final result is obtained. The variables wae and wlstm correspond to the weights given to the normalized Conv-AE and Conv-LSTM scores respectively. The above figure was made with the help of [8] and [11]

.

STM models are trained on datasets that consist of individual frames of non-anomalous videos. By doing this, our model gains the ability to very accurately reconstruct and predict non-anomalous frames with a low loss. However, when an anomalous event or action is encountered, our model does not reconstruct or predict the frame as well, resulting in a relatively higher loss value. We make use of this difference in loss values for anomalous and non-anomalous frames to classify a particular frame as anomalous or non-anomalous. This method (using an Autoencoder model and a ConvLSTM model in parallel), to the best of our knowledge, is a novel approach to detecting anomalies in videos.

3.1 Data

The input to the Autoencoder model is an RGB frame, which is extracted from the video to be classified as anomalous or non-anomalous. All the frames are extracted from the video and undergo some preprocessing. The frames are converted into a pytorch tensor. The frames are resized to have dimensions 128x128. The image pixel values are normalized to have a mean of 0.5 and a standard deviation of 0.25. The pixel values of the image are also clamped between -1 and 1.

The input to the ConvLSTM model is 10 greyscale frames, which are used to predict the 11th frame. All the frames are extracted from the video to be classified. The frames are converted into a pytorch tensor. The RGB frames are converted into greyscale. The frames are resized to have dimensions of 128x128 bits. The pixel values of the image are normalized to have values between 0 and 1 by dividing the tensor by 255.

For both models, the image dimensions are reduced to decrease the time taken for training and testing the model.

3.2 Autoencoder Model

The autoencoder model is used for frame reconstruction. This frame reconstruction using the autoencoder model will help better capture and compare spatial information in the frame, which will help better identify anomalous events. The autoencoder model is composed of two parts used in succession: The encoder and the decoder. The encoder is a deep neural architecture composed of a series of different convolution layers, which are used in succession, to reduce the dimensions of the input image and increase the number of channels of the image. The decoder has the output of the encoder as its input. The decoder has the opposite functionality as that of the encoder. It is composed of a series of different transposed convolution layers in succession, which increase the dimensions of the image and reduce its color channels. The decoder reconstructs the encoder output into the original image having the same dimensions as the encoder input image. This achieves image reconstruction. The input of our autoencoder model is an RGB image of size 128x128 bits and hence has dimensions (3, 128, 128). We have also used batch normalization to help make the training of the model faster and more stable.

(a) Original Image (b) Reconstructed Image

Fig. 3: This is another anomaly example from the CUHK Avenue dataset. The image on the left is the target image and the image on the right is the reconstructed image from our Convolutional Autoencoder. As you can see the man is throwing a bag upwards which is an anomalous action and our autoencoder reconstruction shows significant deformation where the anomalous action is happening.

Our encoder model consists of five 2-dimensional convolution and batch normalization layers with the ReLU activation function. After each convolution

layer, the frame dimensions reduce by a factor of 2 and there is an increase in the number of image channels. The convolution layers also have padding to consider all the pixels of the frame. The output dimensions from the encoder are (256, 1, 1). The image is reduced to a series of 256 pixels.

The decoder model consists of five 2-dimensional transpose convolution layers and batch normalization layers with the ReLU activation function. The last layer of the decoder is followed by the tanh function to get the results between the range of -1 and 1. After each transpose convolution layer, the image dimensions increase by a factor of 2. The output dimensions from the decoder are (3, 128, 128) which is the reconstructed RGB frame.

We have used the ADAM optimizer to train our model with a learning rate of 0.001. The loss function used to compare the original frame with the reconstructed frame is the Mean Squared Error (MSE) loss function. We have trained the autoencoder model on non-anomalous videos. This enables our model to very accurately reconstruct non-anomalous frames with a low loss value. If the frame happens to have an anomalous event in it, the model will not generate an accurate reconstruction of the anomalous frame and the reconstruction will have a relatively higher loss value compared to non-anomalous frame reconstructions, as the model was trained only to reconstruct non-anomalous frames. It cannot accurately reconstruct events that are out of the norm, which in most cases, are anomalies. This higher loss value is compared with a threshold value, which was set after extensive experimentation with different anomalous and non-anomalous videos. If the loss value generated is higher than this threshold value, the frame is classified as an anomaly. This enables our model to identify general anomalies in video using frame reconstruction.

3.3 ConvLSTM Model

We use the ConvLSTM model for future frame prediction. While the autoencoder model is very useful for capturing and understanding spatial information, we also need to capture temporal information as well as detecting anomalies in videos which is a sequence of events occurring with time. The LSTM model helps us achieve this and combat the traditional RNN's vanishing gradient and long-term dependence issue. Since we are dealing with identifying anomalous frames, we also need to capture the spatial information in the image. To achieve this, we have used a modification of the traditional LSTM in which the matrix multiplication operation is replaced with the convolution operation. This is our ConvLSTM model.

$$it = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ei} * C_{t-1} + b_i) \tag{1}$$

$$ft = \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{ef} * C_{t-1} + b_f) \tag{2}$$

$$ot = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} * C_{t-1} + b_o) \tag{3}$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \tag{4}$$

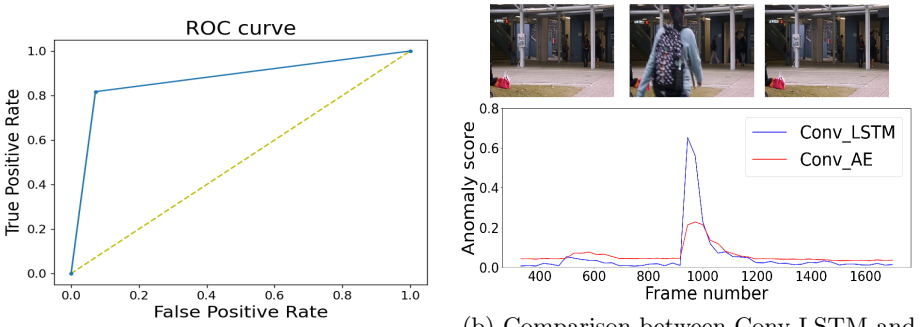
$$H_t = o_t * \tanh(C_t) \tag{5}$$

Equations (1), (2), (3), (4), and (5) represent the cell state and hidden state updation of the ConvLSTM model.

(a) Target Image (b) Predicted Image

Fig. 4: This is an anomaly example from the CUHK Avenue Dataset [12]. The first image shows the target image which our ConvLSTM model is trying to predict, where a man is running. The image on the right shows the frame our model has predicted and it can be seen that there is noticeable deformation since the model is unable to predict the man running, clearly showing the frame is anomalous.

The input to the ConvLSTM model is 10 consecutive grayscale frames, which are resized to 128 x 128 bits. The model uses the spatial information of the images and captures temporal information to output a single frame, which is the predicted 11th frame. Our model is trained using the ADAM optimizer with a learning rate of 0.0001. The predicted 11th frame is compared with the actual 11th frame using the Mean Squared Error (MSE) loss function and a loss value is obtained. The model is trained only on videos that do not have any anomalies. This enables our model to very accurately predict the 11th frame if the previous 10 frames which are given as input do not have any anomalies and the loss value obtained from the MSE function is low. If any of the previous 10 frames which are given as input to the model has an anomaly, the 11th frame is not predicted as well because our model is not trained to predict any anomalous events. This causes the loss value to be relatively higher than non-anomalous frames. This higher loss value enables us to differentiate any general anomalous behavior from non-anomalous behavior. Similar to the autoencoder model, the loss value is compared with a threshold loss value which is set after extensive experimentation with different anomalous and non-anomalous frames. If the loss value is greater than the threshold value, the frame is classified as an anomaly else the frame is not anomalous.



(a) ROC curve for Conv-LSTM.

(b) Comparison between Conv-LSTM and Conv-AE

Fig. 5: Figure (a) shows the ROC curve for the Convolutional LSTM model trained on the Ped2 Dataset[7] with an AUROC score of 87.27%. Figure (b) shows the comparison of the anomaly scores of the Convolutional LSTM model and Convolutional Autoencoder model trained on the CUHK Avenue Dataset [12] with anomaly score on the y-axis and frame number on the x-axis.

4 Experimental Work

4.1 Datasets

The datasets that we have utilized for training and testing our model are the UCSD Ped2 and CUHK Avenue which have been frequently used for video anomaly detection. These datasets are split into training and testing sets where the training set contains only normal scenarios and the testing set contains a mixture of anomalous and non-anomalous frames. Each dataset also contains the ground truth values for each frame in each clip to help train and evaluate the performance of models.

UCSD Ped2: [7] The Ped2 Dataset contains videos of a pedestrian walkway taken from a camera that gives a top-level view of the walkway from a distance. It contains a total of 28 video samples out of which 16 are for training and 12 are for testing and has a pixel resolution of 240x360. The training set contains only normal behavior. The various anomalies that can be found in the testing portion of the dataset are abnormal entities like bikers, small trucks, skaters and abnormal movement of pedestrians into the grass surrounding the walkway.

CUHK Avenue: [12] The Avenue Dataset contains videos of a fixed scene filmed at eye level at the CUHK campus avenue. The Avenue Dataset is divided into 16 training samples and 21 testing samples with a pixel resolution of 360x640. The anomalies are present only in the testing portion of the dataset and include instances such as people throwing various objects(bags and paper) in the air, people moving in the wrong direction or being too close to the camera, and abnormal objects such as bags, cycles, etc.

4.2 Implementation Details

Our implementation includes the Conv-AE, Conv-LSTM, and the combined model. We use PyTorch to implement all our models and the Adam optimizer with a learning rate of 0.0001 for Conv-LSTM and 0.001 for Conv-AE.

For the Conv-AE we resize the input image to 128x128 and normalize the pixel value in the range of [0,1]. We use the Relu activation functions with a Tanh activation function in the final layer.

For the Conv-LSTM we resize the input image to 64x64 with 64 kernels for the Avenue dataset and 128x128 with 128 kernels for the Ped2 dataset, then we convert the image to grayscale and have a kernel size of 3x3. We use the Relu activation functions with a Sigmoid activation function in the final layer.

To combine both the Conv-AE and Conv-LSTM models we collect the frame-wise reconstruction errors and prediction errors from the Conv-AE and Conv-LSTM respectively and store them in separate .npy files. Thus, when the time comes to evaluate the combined model, we load these two .npy files and perform z-score normalization to the two numpy arrays and assign respective weights to the new normalized arrays namely wae and wlstm.

$$zScoreNormalize(x) = \frac{x - \mu}{\sigma} \tag{6}$$

Finally, we combine these two arrays according to the given weights and set a new threshold and compute the combined model AUROC score. By storing and loading the frame-wise reconstruction errors and prediction errors in .npy files it allows us to evaluate the combined model in a fast and efficient manner.

5 Result Analysis

5.1 Evaluation Metrics

Our study follows the popular evaluation metric for Video Anomaly Detection (VAD) which is the Area Under the Receiver Operation Characteristic (AUROC). The reason why AUROC is used as the evaluation metric is because VAD is a binary classification and VAD datasets tend to be highly imbalanced which can be handled by AUROC. The AUROC scores are computed by varying the thresholds for the frame-level anomaly scores produced by the model and comparing these values with respect to the ground-truth annotations. A higher AUROC score indicates better VAD accuracy.

5.2 Main Results

The UCSD Ped2 dataset[7] is considered to be one of the oldest and most commonly used datasets for video anomaly detection. It is considered to be simpler than other VAD datasets such as ShanghaiTech and CUHK Avenue. The CUHK Avenue[12] can be considered to be a step up from the UCSD Ped2 dataset when it comes to complexity.

AUROC SCORES FOR OUR MODELS		
	Avenue	Ped2
Conv-AE	73.47%	68.42%
Conv-LSTM	75.52%	87.27%
Combined Model	77.44%	87.31%

Table 1: The above table presents the respective AUROC scores on the Avenue[12] and Ped2[7] datasets. It can be observed that the combined model has a superior performance on the Ped2 dataset compared to the Avenue dataset which is a common trend that can be seen in other Video Anomaly Detection models. It can also be seen that among the Conv-AE and Conv-LSTM, the Conv-LSTM performs at a relatively higher level compared to the Conv-AE.

Conv-AE Model: The Conv-AE model underperformed when it came to the Ped2 dataset which is why we are working on improvements in that aspect. The Conv-AE model gave an AUROC score of 73.47% on the Avenue dataset with a threshold of 0.1, 80 epochs, and resizing the input frame to 128x128.

Conv-LSTM Model: The Conv-LSTM gives an AUROC score of 87.27% on the Ped2 dataset with a threshold of 10.8, 900 epochs, and resizing the input frame to 128x128. The Conv-LSTM model gives an AUROC score of 75.52% on the Avenue Dataset with a threshold of 33, 100 epochs, and resizing the input frame to 64x64.

Combined Model: Our combined model of Conv-AE and Conv-LSTM gave an accuracy of 87.31% on the Ped2 dataset with a threshold of 0.9 and the weights for the Conv-AE and Conv-LSTM as wae = 0.6 and wlstm = 1 respectively.

Our combined model of Conv-AE and Conv-LSTM gave an accuracy of 77.44% on the Avenue dataset with a threshold of -0.4 and the weights for the Conv-AE and Conv-LSTM as wae = 1 and wlstm = 2.5 respectively. The Conv-AE model was trained for 80 epochs and the Conv-LSTM model was trained for 100 epochs on the Avenue dataset.

Comparative Analysis: The performance of our Conv-AE is comparable to the Conv-AE seen in [5] which has an AUROC score of 70.2% on the Avenue dataset and 90.0% on the Ped2 dataset. The performance of our combined model is comparable to the model seen in [14] which also uses Conv-AE and Conv-LSTM but in a sequential manner and has an AUROC score of 77.0% on the Avenue dataset and 88.1% on the Ped2 dataset. However, our model under performs compared to the state-of-the-art models seen in [11], [9], [4], [10] and so on.

To conclude, although our method does not achieve the performance seen in the state-of-the-art models, it does provide a novel approach that helps capture the temporal capabilities of Conv-LSTM along with the raw anomaly detection capabilities of the Conv-AE and there is a lot of room for improvement such as a memory-based convolutional autoencoder [4] which can help improve the accuracy of our combined model.

6 Conclusion

In this paper, we present a novel approach of parallelizing the results obtained from reconstruction and frame prediction-based methods to detect anomalies in videos. By normalizing the results obtained from both methods, we see that our accuracy improved significantly as compared to using only one by itself. This is because while the autoencoder captures spatial coherence very well, ConvLSTM captures temporal coherence as well, which is important in videos.

7 Future Scope

For future scope, we will explore the combination of memory-augmented autoencoders with ConvLSTM and experiment with different numbers of frames used to predict the target frame in our ConvLSTM model. Enhancing our combined model to a real-time environment is another major improvement that we can work on that can supplement the real-world application of our model.

References

1. Z. K. Abbas and A. A. Al-Ani. A comprehensive review for video anomaly detection on videos. In *2022 International Conference on Computer Science and Software Engineering (CSASE)*, pages 1–1. IEEE, 2022.
2. A. Basharat, A. Gritai, and M. Shah. Learning object motion patterns for anomaly detection and improved object detection. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.
3. M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12742–12752, 2021.
4. D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.
5. M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.
6. P. Y. Ingle and Y.-G. Kim. Real-time abnormal object detection for video surveillance in smart cities. *Sensors*, 22(10):3862, 2022.
7. W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.
8. F. Liu, Z. Zhou, H. Jang, A. Samsonov, G. Zhao, and R. Kijowski. Deep convolutional neural network and 3d deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. *Magnetic resonance in medicine*, 79(4):2379–2391, 2018.
9. W. Liu, H. Chang, B. Ma, S. Shan, and X. Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12147–12156, 2023.

10. W. Liu, W. Luo, D. Lian, and S. Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018.
11. Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13588–13597, 2021.
12. C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013.
13. Y. Lu, K. M. Kumar, S. shahabeddin Nabavi, and Y. Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.
14. W. Luo, W. Liu, and S. Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International conference on multimedia and expo (ICME)*, pages 439–444. IEEE, 2017.
15. J. R. Medel and A. Savakis. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*, 2016.
16. A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab. Machine learning for anomaly detection: A systematic review. *Ieee Access*, 9:78658–78700, 2021.
17. T.-N. Nguyen and J. Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1273–1283, 2019.
18. T. Reiss and Y. Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv preprint arXiv:2212.00789*, 2022.
19. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.
20. W. Sultani, C. Chen, and M. Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018.
21. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
22. D. Xu, Y. Yan, E. Ricci, and N. Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017.