Unsupervised Machine Learning to Detect Abnormal Activities using CNN & 3D Spatial-Temporal AutoEncoder (3DSTAE)

Mandakini Ingle*, Pinky Rane, Harshraj Sharma, Ishika Goyal, Kapil Nakum

Department of Computer Science and Engineering, Medi-Caps University, A.B. Road, Rau, Indore – 453331

*mandakini.ingle@medicaps.ac.in

* Corresponding author

doi: https://doi.org/10.21467/proceedings.7.6.39

Abstract

Live video feeds are very much important now for security, traffic, and keeping factories in check. Deep learning, using things like CNNs and Autoencoders, is really useful for keeping an eye on these videos. This paper talks about using CNNs and Autoencoders to watch live video and catch anything weird. We are teaching a CNN to focus on what matters, and then using an Autoencoder to spot things that do not seem right. The CNN learns what's normal by watching lots of regular video clips. The Autoencoder gets very well at copying those normal clips. When we test it, the Autoencoder looks for trouble by comparing its rebuilt clips with the real ones. The research has been carried out with the UCSD Pedestrian dataset, which has tons of walking scenes. The results say our system is spot-on and better than other ways of finding odd stuff in live video. This could be a game-changer for security, traffic, and factories where you need to catch problems ASAP. So, this study says that CNNs and Autoencoders are a good team for watching video and finding weird action as it happens. It also says that deep learning can really help with video checking in all sorts of places.

Keywords: camera, long term care, autoencoder, deep learning, computer vision.

1. Introduction

Security's a real pain these days with crime on the rise. Live video and quick reactions can really stop criminals and keep things safe. But watching a bunch of screens all day is tough. People get tired, miss stuff, or make mistakes.These days, people are using smart technology and machine learning to help watch video feeds for anything unusual. Convolutional Neural Networks (CNNs) and autoencoders work really very nice for this [1]. CNNs are good at clicking out the important parts of a video and learning what normal looks like [2]. Autoencoders, on the other side, work by simplifying the video to its core elements and then reconstructing it. By adding CNNs and autoencoders, the system can capture strange or unexpected events more faster. Together, they learn what "normal" looks like, and when something out of the ordinary happens, the system can alert security right away. In this research paper, we're going to chat about a cool way to watch videos and catch unusual stuff using CNNs and autoencoders [3]. Our system has three steps: (1) grabbing the video, (2) picking out the key things, and (3) spotting anything strange and sending out alerts [4]. The video comes from cameras and the system preps it for the computers. Then it pulls out the important info and makes the video smaller. After all that, the system searches for anything strange in that smaller video and then tells security if needed. Fig. 1 explains about how the video is captured where CNN extracts the important features from each video frame and the autoencoder reconstructs the feature sequence.

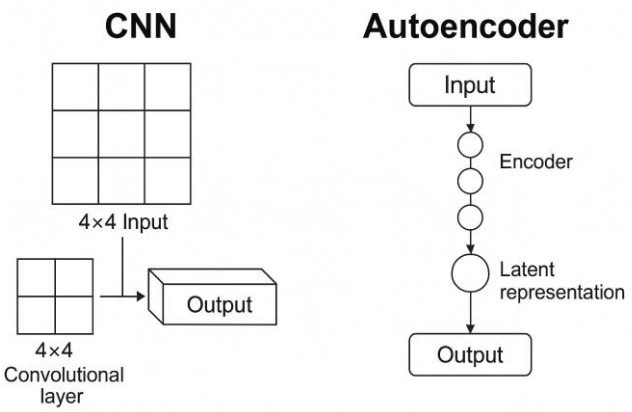

Fig. 1. Architecture Design of CNN and autoencoder

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

To see if the system is worth it, the research has been tested using a video collection that's open to the public. The research stacked our system up against others. The tests showed that our setup is better at spotting stuff and doesn't mess up as much. Here's the overview of this research paper- The research take a look at what others have done before us in terms of monitoring video and detecting unusual activity. The research dive into how our system works, explaining how we set up and train the CNNs and autoencoders to do their job. The research talk about the tests and what we learned. The research figure out what the test results mean.The research wrap up and talk about where things could go from here.Using CNNs and autoencoders to check videos and quickly find unusual events is a good way to help keep places safe. Our idea shows that these systems can be useful, and using computer tricks in video setups is smart.

2. Literature Review

The demand for surveillance and security measures has increased, leading to a growing interest in video- based crowd behaviour analysis and anomaly detection [5]. Computer vision technologies have emerged and researchers are exploring new ways to extract statistical features from video footages for real-time detection of abnormal behaviours. This literature review looks at recent research on detecting unusual behaviour in crowd scenes, focusing on using spatiotemporal textures (features that track both space and time) for anomaly detection [6]. It discusses the strengths and weaknesses of current methods and points out the main challenges and areas for improvement in future research. This research paper introduces a new way to spot unusual stuff in videos using something called the Incremental Spatio-Temporal Learner (ISTL) [7]. It's all about using deep learning to watch videos in real-time. It deals with the difficulties of processing lots of video data, figuring out what's normal, and keeping up with changes in what's normal using some smart tricks. The ISTL uses a special model made of layers that can spot patterns in space and time . It also uses two limits to avoid missing things. Tests on datasets show it's accurate, solid, not too hard on computers, and works in different places like factories and cities. The results back up that it's good for watching videos in real-time. The authors plan to make a fully automatic video-watching system that doesn't miss as many bad things by using a complicated system that learns and adapts. Basically, this paper has a cool new idea that could be a big deal for video surveillance. Recent techniques work on spatiotemporal learning. For example, MTFL(Multi Time Scale Feature Learning) uses a video transformer to effectively capture short, medium and long term pattern in surveillance footage, achieving state of the art results on datasets [8]. This research paper talks about a fresh method for spotting strange things in videos by using training data that's not label perfectly. The method gets better results than other methods. This is shown in a big dataset of real surveillance videos with weird and normal stuff. The paper also gives some basic results to start with for recognizing strange activities, which shows that the dataset is hard and could be used for more study. All in all, this paper is a helpful addition to the field of video anomaly detection. A Two Stream Convolutional Network integrated with a Multiple Instance Learning (MIL) framework has formulated the anomaly detection possibility with limited supervision [9]. This helps to reduce the efforts of dependence on frame- level annotations. Similarly, STPrompt also uses spatio-temporal prompts from pre-defined language models that helps to localize anomalies without explicit supervision [10]. This research paper suggests a way to automatically spot events that shouldn't be happening using images from cameras. It takes some simple features from different parts of the picture and spots strange events by checking how the features change when the picture quality drops or the view changes. The decision-making method uses a simple way to add up two growth factors and a filter. Tests show the algorithm is accurate and doesn't give many false alarms in real situations. The paper explains the method and results well. Basically, it's a good step forward for spotting camera problems. The concept of lightweight processing also displays in newer hybrid models. For example, a method combines YOLOv7 for spatial object detection with RNNs for temporal sequence structure that provides fast and effective real-time video anomaly detection [11]. This research paper goes over previous work, like using real-time security cameras and machine learning to spot weird behaviour in crowds to make things safer [12]. Methods that use deep learning and combinations of methods work the best. But some methods can be too slow. The authors want to try out some new things to make it more accurate and faster. This paper gives a good summary of how things are done right now for spotting weird stuff in crowds. This research paper talks about a real-time method that doesn't need supervision for spotting strange things in video surveillance. It uses features that capture what things look like and how they move in a scene. The method uses a simple way to model things and look at local patterns to spot anomalies. It also updates the scene to handle changes. It was tested on datasets and did better than other methods at spotting different types of anomalies, especially in crowded scenes where things don't move as expected. This paper presents a good method for improving anomaly detection in video surveillance for security reasons. This

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

research paper suggests a way to automatically find objects and group their paths to spot anomalies in video streams in real-time. The method is good at spotting new things and can handle noisy data. The results show it's faster than other methods. The authors want to work on dealing with moving cameras. This paper provides a better approach for anomaly detection. Based upon this idea, MissionGNN introduces a graph-based model that focuses on high-level knowledge that help us to recognize anomalies with minimal human supervision and adapt to environment condition [13]. This research paper suggests a new method that concatenate putting things back together and guessing future frames to spot anomalies in videos [14]. The method uses a network with two blocks to predict and rebuild things [15]. This gives better results. The method is good at filtering out noise and works in real-time. All in all, this paper has a good solution.The authors talk about a real-time video surveillance system that uses learning to make a model better at spotting anomalies [16]. They tested it and shows that it can spot anomalies in real-time without too many false alarms.

3. Methodology

The research has been done to make a system that could watch live video, analyze it and see when something weird happens. The idea uses both a CNN and a 3D STAE. Here's how it goes: Data Time: First of all, a lot of video has been needed to teach the system. So, a dataset with videos of places such as parking lots, streets, and shopping spots is required. These videos had both normal stuff, and unusual stuff too. This helped to test if the system could really find anomalies. We used the USCD dataset [17]. It has two parts: Ped1 (34 training videos and 36 test videos at 238 × 158) and Ped2 (16 training videos, 12 test videos at 360 × 240) [18]. The camera angles are not the same between Ped1 and Ped2. Fig.2 show the examples of different anomalies (Tang et al., 2020b) [19].

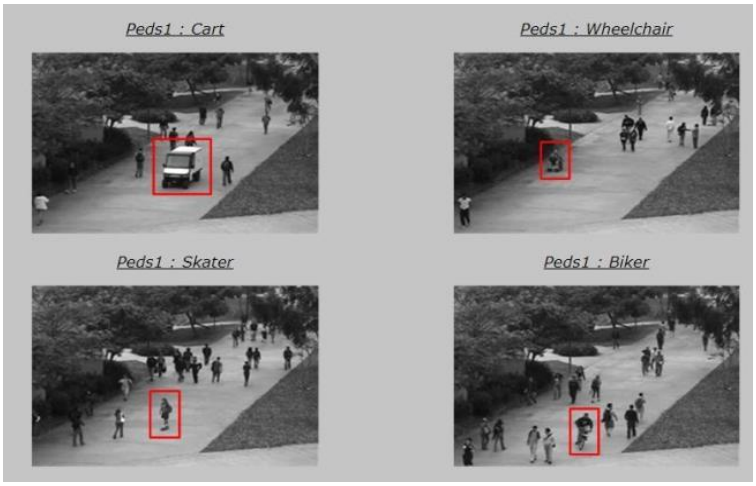

Fig. 2. Example videos

The above figure shows the unusual instances of cars, skaters and wheelchairs that appear in pedestrian areas. Anomalies are things such as bikes, cars or wheelchairs in the places where they shouldn't be present there. Getting the Data Ready:Before using the videos, it needs to be prepared. This meant loading each video frame, making it into numbers that the network could read, and resizing it to (227,227,3). Every frame was converted to black and white. To give the training data a boost, Keras ImageDataGenerator has been used to change the videos a bit – like rotating, shifting, and changing the colors. This helped in training the model. Model Deets: CNN with a 3D STAE is used. The CNN grabs details from each video frame by using two layers (128 and 64 filters), plus a max pooling layer (pool size: (2, 2), stride: (2, 2)). The tanh thing is used after each layer to add action and a batch thing to help the network learn faster. The STAE is used for watching live and spotting anomalies. Fig. 3 shows the 3D STAE design that illustrates how the input frame goes through convolutional layers, LSTM layers and deconvolutional layers (Khan et al., 2022) [20].

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

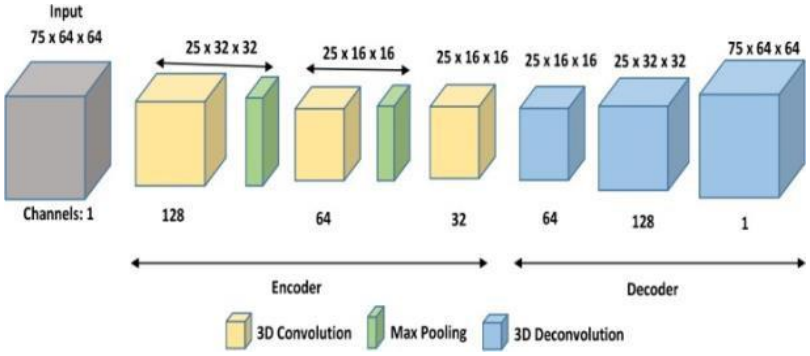

Fig.3. 3DSTAE Spotting Strange Behavior in Video

Using the Keras API, it depends on layers to learn movement in videos [1].The STAE model tries to catch odd stuff that doesn't match normal video. Tanh activation is used to improve model dynamics. 3D STAE process the extract features to find abnormal activity. It learns normal by training on videos of normal stuff. Later, it can watch live video and point out weird things. Table 1 lists the stuff in the STAE model. It has seven layers: three convolutional layers , three LSTM convolutional layers, and one deconvolutional layer. The first layer has 128 number of filters, a kernel size of (11,11,1), a stride of (4,4,1), and the tanh activation. It learns details in the video frames. The convolutional and LSTM layers get the picture and watch movement in the videos. The STAE model is trained to recreate the input using mean squared error thing and the Adam stuff, including the features it has learned. Accuracy is measured by checking the mean squared error between what it predicted and what the frames actually were. The STAE model appears useful for watching live and for catching anomalies. It gets normal movement and flags the weird stuff. The sound and the depth of objects in the frame can be added to help it.

Table 1. Model Layer

Layer Type	Output Shape	Number of Parameters
Input Layer	(227,227,10,1)	0
Conv3D Layer	(55, 55, 10,128)	15,616
Conv3D Layer	(26, 26, 10, 64)	204,864
ConvLSTM2D Layer	(26, 26, 10, 64)	199,680
ConvLSTM2D Layer	(26, 26, 10, 32)	110,720
ConvLSTM2D Layer	(26, 26, 10, 64)	442,624
Conv3DTranspose	(55, 55, 10,128)	204,928
Conv3DTranspose	(227, 227, 10,1)	15,617
Total		1,193,049

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

4. Algorithm

Identify and mark the content of the video to understand whether the object is normal or if some kind of abnormal behavior is detected to understand whether something unusual exists or not. Get the videos ready and extract each frame from the videos.Turn each frame into a small image (227x227 pixels). Turn the photos to black and white so your computer does not have to work as hard.Get creative with the photos a little to enhance the model.Rotate them slightly.Shift them left or right.Shift them up or down.Slightly change the colors.Normalize the color values to values between 0 and 1.Bunch the frames into brief, 5-second clips (75 frames each).

Build the Brains of the System. Part 1: CNN for Perceiving things [21].Use a CNN with:One with 128 filters, a bigger kernel, a stride of (4, 4, 1), and tanh activation. Another with 64 filters but a smaller kernel. Use Max Pooling (2,2) after each filter.*Use batch normalization to facilitate training of the model.* Part 2: 3D Spatial-Temporal Autoencoder (STAE) for Spotting Anomalies: *It is made up of convolutional and ConvLSTM layers:* Conv3D Layer (128 filters) *Conv3D Layer (64 filters)* ConvLSTM2D Layer (64 filters) *ConvLSTM2D Layer (32 filters)* ConvLSTM2D Layer (64 filters) *Conv3DTranspose Layer (128 filters)* Conv3DTranspose Layer (output layer) *The goal is to make the output look like the input.*Use Mean Squared Error (MSE) to score how well the model is performing.

Train the Model.Train the CNN initially in order to get it to learn what to pay attention to in the leading areas of the video.Train the STAE on normal video clips second.*Observe how well the model is doing relative to its original input.* Train for 5 cycles. Look for Anomalies in New Videos.Pass the test video through the model you've trained.Get the CNN and STAE to pull out the key features.Look at how much the model's output differs from the input.If the difference is more than some threshold, call it weird.Trigger an alert if something weird is found. Observe How Well It Works. Try how well it works in comparison to other basic methods (such as optical flow or motion detection).Look for these things: Total correctness and how well it is when it detects something. Apply the algorithm and Run the model using Python 3.9 and OpenCV.Set the system up with a webcam to observe live video. Save the video (at 128x128 pixels and 15 frames per second) for future verification.

Data Collection- Surveillance videos are gathered for training and testing. They are employed to identify the normal and anomalous (abnormal) behaviors. The data may be real-world environments such as parking lots, roads, or malls (e.g., the UCSD Pedestrian Dataset) [18]. Preprocessing and data augmentation- Frames are extracted from the video, resized, converted to grayscale, and normalized. Augmentation techniques like rotation and shifting are carried out using libraries like Keras' ImageDataGenerator to improve model generalization and avoid overfitting.

Feature extraction with CNN- a convolutional neural network is applied to extract the spatial features of every frame which will aid in decreasing the video data complexity by converting it to quality feature maps. Training 3D Spatial-Temporal Autoencoder (STAE)-These features are input to a 3D STAE, which learns the spatial-temporal behaviors of normality in the video. Its encoder-decoder model enables it to reconstruct normal sequence videos. Model Evaluation and Testing- The model is tested against a test set in order to analyze its ability of detecting anomalies. Accuracy, false positives/negatives, and mean squared error are calculated. Real-Time Anomaly Detection- The trained model is then utilized in real time to analyze incoming streams of video. Anomalies are identified by contrasting reconstructed frames with real frames.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

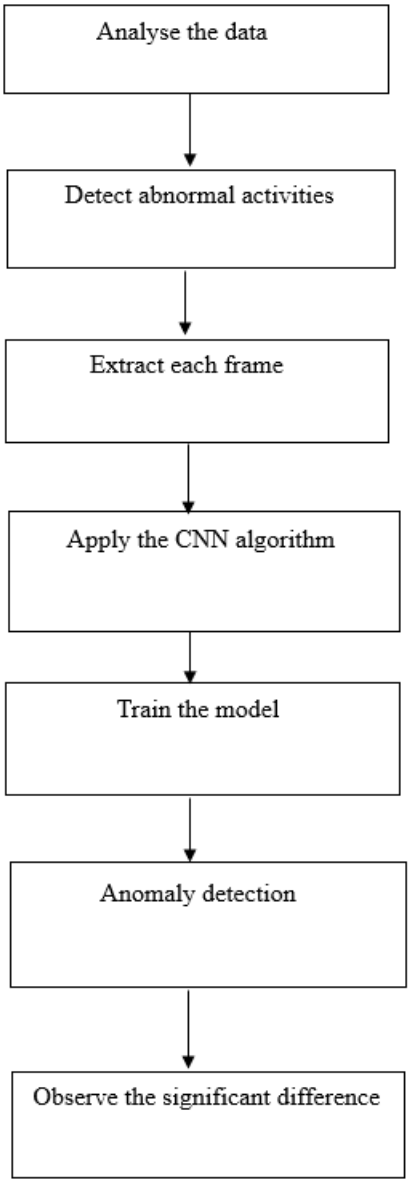

Fig. 4. Workflow of anomaly detection system

5. Result and Analysis

The research consists of a live video surveillance system that uses a mix of CNN and Autoencoder to spot odd stuff. The system has been trained and tested using the security videos from different places and checked how well it worked using various measurements.The CNN and Autoencoder system was been trained with a bunch of video clips. After training it five times, it got about 75% correct, with a loss of 0.07%. The training results tell us that the system learned to pretty accurately pick out unusual things in live video. This system can be used to detect crimes, weird actions in public, and possible security problems. The research has been tested everything out using VS-Code Editor on Windows. The code has been written in Python 3.9 to train the system. Then, the software has been linked to a webcam using openCV to get live video. The system can handle videos that are 128x128 pixels at 15 frames per second. CNNs and Autoencoders are used in our system to get good accuracy. CNNs are great for looking at images and videos because they can pull out key parts of the visual data. Autoencoders help shrink things down and grab important features. By putting these together, the system can find small problems in the video that might be missed by regular methods. Afterall, this study shows that use of CNN & Autoencoder for real time video security is the best way to find unusual stuff in real-time. Still, the things can be tweak to make it even better, like adjusting the model's settings and trying out other ways to spot anomalies. So, this system could be a helpful tool for making things safer in lots of situations. More work in this area could lead to even better live security systems.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

The model proposed with Convolutional Neural Networks (CNN) for feature extraction and a 3D Spatial-Temporal Autoencoder (STAE) for anomaly detection had the best accuracy of 75.48% in the comparison. Optical flow, a traditional algorithm that prioritizes motion across frames, ranked at lower accuracy (~68.5%), tends to fall in noisy and dynamic scenes [22]. Motion detection , another classical method based on frame differencing, performed slightly better [23] than optical flow (~70.3%) but was still behind the deep learning model. Fig. 5 shows the readily contrast of the superiority of the deep learning-based approach proposed.

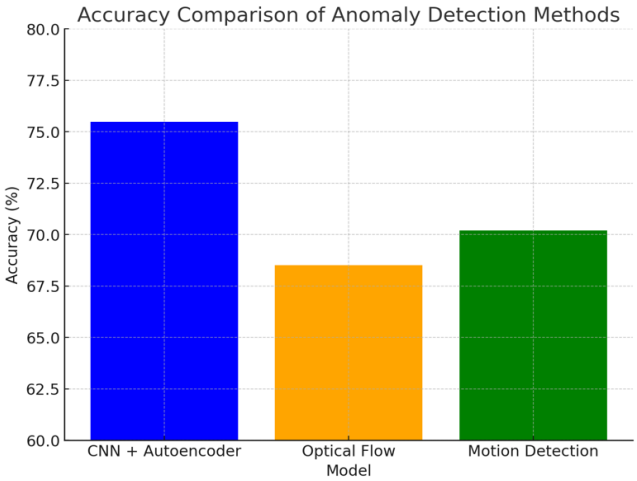

Fig. 5. Model Accuracy Analysis

An accuracy comparison was performed among the proposed CNN + Autoencoder model, Optical flow model and traditional motion detection methods. The CNN + Autoencoder shows the highest accuracy. The spatial feature extraction capability of CNNs and reconstruction along with anomaly detection of autoencoders makes them very effective at spotting such subtle abnormal behavior.

6. Conclusion

In this study, a proposed method for real-time video surveillance and anomaly detection that makes use of CNN and autoencoders is discussed and evaluated. The process comprises autoencoding for anomaly detection and training a CNN for feature extraction. The autoencoder is trained to mimic the normal frames, whereas the CNN algorithm is trained to learn the usual patterns of the scene using a huge dataset of normal video frames. By comparing the reconstructed frames to the input frames, the autoencoder is employed in the evaluation step to detect abnormalities. The approach was tested on the Avenue dataset, containing a range of pedestrian activity situations. The experimental results show that the proposed approach performs better than current methods for real-time video abnormality detection and reaches high accuracy.

References

[1] Du Tran, Rainer Sorokin, Gerard Medioni. Long short-term memory over observation times for activity recognition. European Conference on Computer Vision (ECCV). https://doi.org/10.1007/978-3-319-10599-4_18
[2] Nawaratne, R., Alahakoon, D., De Silva, D., & Yu, X. (2020). Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance. IEEE Transactions on Industrial Informatics, 16(1), 393-402. https://doi.org/10.1109/tii.2019.2938527
[3] Nguyen, H., Loan, T.T.K., Mao, B. D., & Huh, E (2015). Low cost real-time system monitoring using Raspberry Pi. In International Conference on Ubiquitous and Future Networks. https://doi.org/10.1109/icufn.2015.7182665
[4] Kim, J., & Grauman, K. (2009). Observe Locally, Infer Globally: A Space-Time MRF for Detecting Abnormal activities with Incremental Updates. https://doi.org/10.1109/CVPR.2009.5206757
[5] Ko, T. H. (2008). A survey on behavior analysis in video surveillance for homeland security applications. In Applied Imagery Pattern Recognition Workshop. https://doi.org/10.1109/aipr.2008.4906450
[6] Xu, D., Ricci, E., Yan, Y., Song, J., & Sebe, N. (2015). Learning Deep Representations of Appearance and Motion for Anomalous Event Detection. https://doi.org/10.5244/C.29.8
[7] Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning Temporal Regularity in Video Sequences. https://doi.org/10.1109/CVPR.2016.597
[8] Zhang et al., 2024. https://doi.org/10.48550/arXiv.2410.05900
[9] Nejad & Haque, 2024. https://doi.org/10.48550/arXiv.2411.08755
[10] Wu et al., 2024. https://doi.org/10.48550/arXiv.2408.05905
[11] Poirier, 2024. https://doi.org/10.48550/arXiv.2410.15909

[12] Rezaee, K., Rezakhani, S. M., Khosravi, M. R., & Moghimi, M. M. (2021). A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing. https://doi.org/10.1007/s00779-021-01586-5
[13] Yun et al., 2024. https://doi.org/10.48550/arXiv.2406.18815
[14] Liu, W., Luo, W., Lian, D., & Gao, S. (2018). Future Frame Prediction for Anomaly Detection – A New Baseline. https://doi.org/10.1109/CVPR.2018.00957
[15] Tang, Y. L., Zhao, L., Zhang, S., Gong, C., Li, G., & Yang, J. (2020). Integrating prediction and reconstruction for anomaly detection. Pattern Recognition Letters, 129, 123-130. https://doi.org/10.1016/j.patrec.2019.11.024
[16] Sadeghi-Tehran, P., & Angelov, P. (2012). A real-time approach for novelty detection and trajectories analysis for anomaly recognition in video surveillance systems. In 2012 IEEE Conference on Evolving and Adaptive Intelligent Systems. https://doi.org/10.1109/eais.2012.6232814
[17] Wang, J., & Xu, Z. (2016). Spatio-temporal texture modelling for real-time crowd anomaly detection. Computer Vision and Image Understanding, 144, 177-187. https://doi.org/10.1016/j.cviu.2015.08.010
[18] https://paperswithcode.com/dataset/ucsd
[19] Tang, Y. L., Zhao, L., Zhang, S., Gong, C., Li, G., & Yang, J. (2020). Integrating prediction and reconstruction for anomaly detection. Pattern Recognition Letters, 129,123-130. https://doi.org/10.1016/j.patrec.2019.11.024
[20] Khan, S. S., Mishra, P. K., Javed, N., Ye, B., Newman, K., Mihailidis, A., & Iaboni, A. (2022). Unsupervised Deep Learning to Detect Agitation From Videos in People With Dementia. IEEE Access, 10, 10349-10358. https://doi.org/10.1109/access.2022.3143990
[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention augmented convolutional networks. https://doi.org/10.1109/ICCV.2017.488
[22] Wang, Y., Fan, C., Cheng, K., & Deng, P.S. (2011).
Real-time camera anomaly detection for real-world video surveillance. In International Conference on Machine Learning and Cybernetics. https://doi.org/10.1109/icmlc.2011.6017032
[23] Bertini, M., Del Bimbo, A., & Seidenari, L. (2012). Multi-scale and real-time non-parametric approach for anomaly detection and localization. Computer Vision and Image Understanding, 116(3), 320-329. https://doi.org/10.1016/j.cviu.2011.09.009