# Predictive Convolutional Long Short-Term Memory Network for Detecting Anomalies in Smart Surveillance

[1]Priyanka Patel, [2]Dr. Amit Nayak

•

[1]U & P U Patel Department of Computer Engineering,
Chandubhai S Patel Institute of Technology,
Faculty of Technology & Engineering,
Charotar University of Science and Technology (CHARUSAT), Changa, Gujarat, India
priyankapatel.it@charusat.ac.in

[2]Department of Information Technology,
Devang Patel Institute of Advance Technology and Research (DEPSTAR),
Faculty of Technology& Engineering, Charotar University of Science
and Technology (CHARUSAT), Changa, Gujarat, India
amitnayak.it@charusat.ac.in

**Abstract**

*Surveillance is the monitoring of behavior, actions, or information, with the purpose of collecting, influencing, controlling, or guiding evidence. Despite the technical traits of cutting-edge science, it is difficult to detect abnormal events in the surveillance video and requires exhaustive human efforts. Anomalous events in the video remain a challenge due to the occlusions of objects, different densities of the crowd, cluttered backgrounds & objects, and movements in complex scenes and situations. In this paper, we propose a new model called time distributed convolutional neural network long short-term memory Spatiotemporal Autoencoder (TDSTConvLSTM), which uses a deep neural network to automatically learn video interpretation. Convolution neural network is used to extract visual features from spatial and time distributed LSTM use for sequence learning in temporal dimensions. Since most anomaly detection data sets are restricted to appearance anomalies or unusual motion. There are some anomaly detection data-sets available such as the UCSD Pedestrian dataset, CUHK Avenue, Subway entry-exit, ShanghaiTech, street scene, UCF-crime, etc. with varieties of anomaly classes. To narrow down the variations, this system can detect cyclists, bikers, skaters, cars, trucks, tempo, tractors, wheelchairs, and walkers who are walking on loan (off the road) which are visible under normal conditions and have a great impact on the safety of pedestrians. The Time distributed ConvLSTM has been trained with a normal video frame sequence belonging to these mentioned classes. The experiments are performed on the mentioned architecture and with benchmark data sets UCSD PED1, UCSD PED2, CUHK Avenue, and ShanghaiTech. The Pattern to catch anomalies from video involves the extraction of both spatial and temporal features. The growing interest in deep learning approaches to video surveillance raises concerns about the accuracy and efficiency of neural networks. The time distributed ConvLSTM model is good compared to benchmark models.*

**Keywords:** Predictive ConvolutionLSTM, Distributed convolutional neural network, long short-term memory, Spatiotemporal Autoencoder, TDSTConvLSTM, Video Surveillance, Anomaly detection, Machine Learning, Deep Learning, Smart Video Monitoring.

# 1. Introduction

Anomalies are patterns or observations in the video sequence that do not conform to normal or expected behavior. And the detection of anomaly is an approach to discover the patterns those unexpected behavior. The unexpected behaviors are characterized as outliers, novel or anomalies. Mainly three types of anomalies are there. First is point anomaly, second is contextual anomaly and the third is collective anomaly. The most common type of anomaly is point anomaly and has been a focus of most of the research. It can be defined as an individual entity that is considered abnormal with regard to other data. In Figure-1 point A1, point A2, and points in region A3 are point anomalies as they are outside to the normal region. The contextual anomaly is also known as a conditional anomaly. It can be defined as in some specific context if a data instance is anomalous then it is a contextual anomaly. These types of anomalies are usually found out in spatial data and time-series data. The last one is the collective anomaly; it can be defined as a set of related data instances that is anomalous with regard to the remaining data in the data set. Individual entities may not be anomalies by themselves in collective anomalies but their occurrence to gather is considered anomalous. The aspect of anomaly detection approach, the nature of data and the type of label on instance are shown in [1, 2]. In recent years, the demand for visual surveillance has increased. Large-scale visual surveillance system has been implemented with many high-quality cameras for security concern in private and public zones but at same time it generates large amounts of data every second and which is impossible to monitor and process such a huge amount of information in real-world applications. Therefore, it is commanding to develop autonomous systems that can detect, identify, and predict abnormal articles or events, and then help take early action to avoid threats or unexpected actions. The international border security agencies are also looking for the solution of security in surveillance in each sector. It can also be used in various applications, such as group activity detection, home security, organization Security, restricted area monitoring, traffic analysis, and also Through macro applications (traffic surveillance, building surveillance, city surveillance, and business intelligence), through micro applications (perimeter intrusion detection, pattern recognition, people counting and management, automatic license plate recognition (ALPR), incident detection, face recognition and others including advanced eye recognition and behavior analysis), upon deployment (on premise and in the cloud), by industry (government and transport, BFSI, trade and industry, healthcare, retail and others, including education, housing and hospitality), by region (North, East, West AND South) and competitive landscape covered by Dublin, Nov. 19, 2018 Research Markets[43,49]. Recent research shows that it has received great attention in the research community and has become a major problem to find a better solution in computer vision. However, the implementation of surveillance systems in practical applications brings three main challenges: one is the unlabeled data that is readily available, but training on the labelled data is not offered. The second is Anomalies that are not clearly defined in actual video surveillance. And the third is Video complexity that adds expensive features and makes it difficult

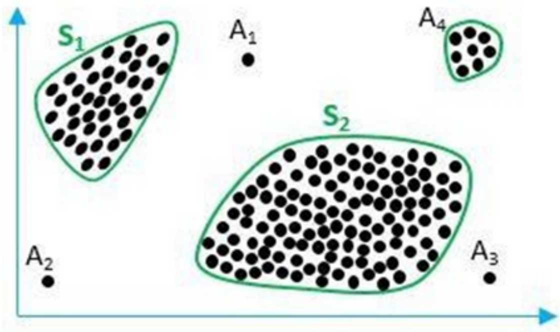



**Figure 1:** *Example with Normal and anomaly region in Two- Dimensional Data*

to extract manually. According to the exclusive summary by global endowment of international Peace [3]. It is important that AI surveillance is not a stand-alone tool of repression, but part of a set of digital tools of repression: information and communication technologies used to monitor, intimidate, coerce and harass people, and to deter certain people Illegal activities or beliefs. A lot of work has been done in the field of video surveillance, such as detecting objects, tracking them, and recognizing the behavior of objects, but it is still interesting to see rare, novel and unusual objects. The interesting thing to occurrences of the new objects in the video frames and at a same time it is also difficult to spot and detect those new and suspicious behavior in large amount of data sequence. Finding such rare occurrences in a video sequence is a critical task for an autonomous model because the unusual and novel, doesn't keep happening again and again, the model has not been adequately trained with these novel objects (anomalous objects). However, current technology works well [4,5,43,49,78] it gives good results, but at the same time they are all contextual independent. The detection of all types of anomalous objects in all data contexts has yet to be introduced. In the case of video data, this is a challenge due to the high dimensional of input data, the noise it contains and the large number of new types of objects and interactions. These objects are context-independent. For example: walking in a canteen is considered an unusual occurrence, but walking in a playground would be normal.

## 2. Background

Illustrates anomalies in a simple two-dimensional data-set is shown in Figure-1. In the whole data set, there are two normal regions which are S1 and S2 since most observations lie in these two regions. Points A1, A2 and A3 and A4 are far away from the normal regions S1 and S2. So those A1, A2, A3 & A4 regions are consider as an anomaly as it does not lie under S1 and S2 normal regions. Various examination studies have examined Decision Engine approaches which can be classified in five categories, classification based, Clustering based, knowledge based, combination based and statistical based, as illustrated in Figure-2.

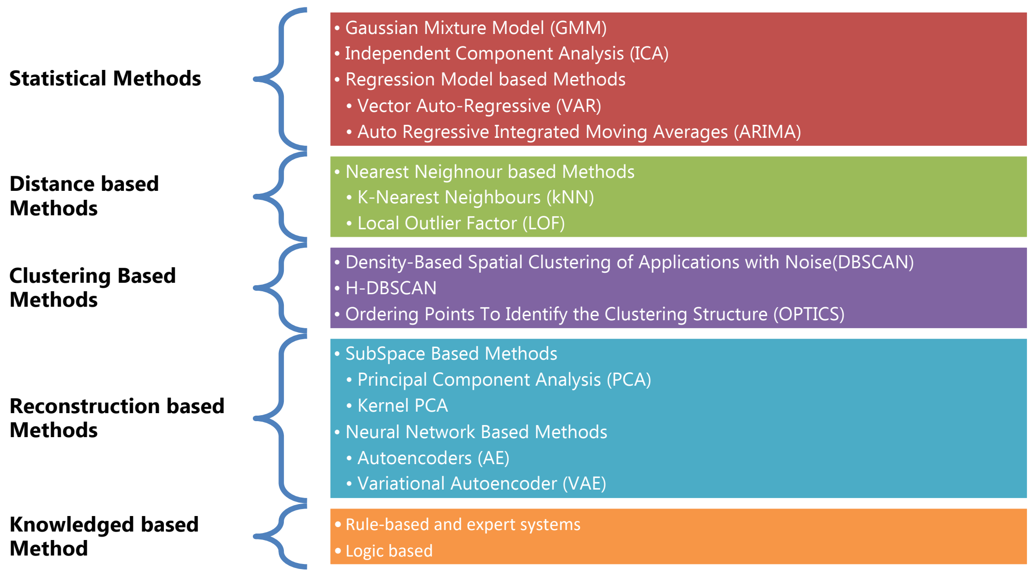



**Figure 2:** *Taxonomy of Classical Methods in Previous Examinations*

The surveillance may include the use of electronic devices such as closed-circuit television (CCTV) for remote monitoring or the interception of Internet traffic and other electronic transmissions of information. It can also include simple technical methods, such as collecting information and intercepting messages. Object Recognition, Detection, and Localization in Video are broadly discovered by researchers in computer vision and machine learning. Figure-3 shows summarize each technique and its corresponding global implementation level [3]. However, it is a monotonous task to spot a particular object or action which is not expected to be a part of the video sequence is relatively difficult to track. Anomalies in surveillance video are extensively described as activities or sports which might be uncommon and symbolize abnormal behavior. It is an unsupervised method of detecting anomalies. Due to the unclear definition of video events, it is challenging to automatically detect abnormal events in long video sequences.

In computer vision, the detection of anomaly events is one of the most difficult problems and has attracted a great deal of research effort in the last few decades [2, 5-8, 49], whereby common detection methods can be roughly divided into the following three groups. The first category of abnormality detection methods focuses on the hypothesis that abnormalities are rare and that behaviors that differ from normal patterns are considered abnormal. In these methods, regular patterns are coded by various statistical models, z field-based models that combine dynamic models and treat anomalies as outliers. The second category of anomaly detection approaches is sparse reconstruction [12] which is used to learn common patterns. In particular, a dictionary is created using a sparse representation for normal behavior, and those with a high error are recognized as anomalies. Recently, with the promising advancement of deep learning, some researchers are building deep neural networks for anomaly detection, including learning video prediction and learning abstraction features [2, 13-15], The third group is the Hybrid methods of normal and abnormal behavior for modeling [10, 17], in which multi-instance learning (MIL) is used in a weakly monitored environment to model movement patterns [15, 17]. For example, Sultani et al. developed a classifier based on MIL [10] that can detect anomalies. In the meantime, a deep classification model is used to predict anomaly scores. To take advantage of the superiority of Sultan's work, which takes normal and anomalous video into account, in this work reconstruct of the model using weakly labeled supervised learning. The related methods based on autoencoder for abnormality detection are describe in [4, 14, 26-28].

In order to avoid enormous computational effort caused by backtracking in the existing methods, we propose a new and effective three-step method for unsupervised anomaly detection.

| AI Surveillance Technique | Description | Global Proliferation (out of 75 countries) |
|---|---|---|
| Smart Cities/Safe Cities | Cities with sensors that transmit real-time data to facilitate service delivery, city management, and public safety. Often referred to as "safe cities," they incorporate sensors, facial recognition cameras, and police body cameras connected to intelligent command centers to prevent crime, ensure public safety, and respond to emergencies. Only platforms with a clear public safety focus are incorporated in the index. | 56 countries |
| Facial Recognition Systems | Biometric technology that uses cameras (still images or video) to match stored or live footage of individuals with images from databases. Not all systems focus on database matching; some systems assess aggregate demographic trends or conduct broader sentiment analysis via facial recognition crowd scanning. | 64 countries |
| Smart Policing | Data-driven analytic technology used to facilitate investigations and police response; some systems incorporate algorithmic analysis to make predictions about | 53 countries |

**Figure 3:** *Summary of AI Surveillance Techniques and Global Prevalence [3]*

In the first stage, a Convneted auto- encoder is trained to learn the latent space representation of frame sequence. Latent space representation of each frame is passed to the stacked thereof or stacked LSTM and it's trained to predict whether the sequence of frames contains an anomaly or not. We proposed such a problem by learning collecting features from previous feature maps that can recognize, detect and localize anomalies under the video. We approach end-to-end trainable composite Time Distributed Spatiotemporal Convolutional Long Short-Term Memory (TSConvLSTM).

**Basic Convolutional LSTM Architecture:**

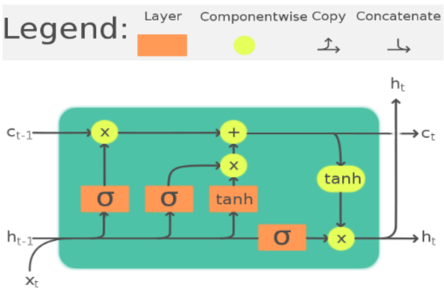



**Figure 4:** *The LSTM cell can process data sequentially and keep its hidden state through time [15].*

In this section, we have explained the types of LSTM networks as shown in Figures 4, 5, 6, 7, 8 respectively and with the help of that we have designed a new hybridTime Distributed ConvLSTM network. LSTM stands for "Long Short Time Memory", it is a layer that can receive various chronological entries in order to find what is very useful for forecasting. It's a simplified explanation, but it's close to reality. Since we process frames in chronological order, we want to be able to determine the relationship from one frame to another at a specific point in time. Since we have timed inputs, LSTM is very suitable for filtering the useful values from these inputs. Usually there are two options: Before applying LSTM, do a convolution or some other neural Computation. Do the same work after the LSTM. To decide which order to choose, you need to think about what you want to filter. In our example, we need to check for a moving object, so we need to find the object before it can detect movement, so here we need to do convolution before LSTM. Here in LSTM input and state at a timestamp are 1D vectors. Dimensions of the state can be permuted without affecting the overall structure. With the increasing popularity of LSTMs, various changes to the traditional LSTM architecture have been tested to simplify the internal design of cells to operate more efficiently and to reduce computational complexity. Gers et.al,2000 introduced peephole connections that allowed the gate layers to see the state of the cell at all times. Some LSTMs also used a coupled entrance and a forgetting gate instead of two separate gates, which helps to make two decisions at the same time.

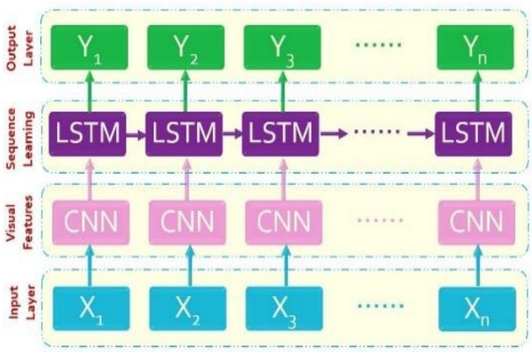



**Figure 5:** *Simple Convolutional LSTM Architecture*

$$i_t = \boldsymbol{\sigma}(W_i \circ x_t + b_{ii} + U_i \circ h_{t-1} + b_{hi}) \tag{1}$$

$$f_t = \boldsymbol{\sigma}(W_f \circ x_t + b_{if} + U_f \circ h_{t-1} + b_{hf}) \tag{2}$$

$$g_t = \boldsymbol{tanh}(W_g \circ x_t + b_{ig} + U_g \circ h_{t-1} + b_{hg}) \tag{3}$$

$$o_t = \boldsymbol{\sigma}(W_o \circ x_t + b_{io} + U_o \circ h_{t-1} + b_{ho}) \tag{4}$$

$$c_t = f_t \circ C_{t-1} + i_t \circ g_t \tag{5}$$

$$h_t = o_t \circ \tanh ct \tag{6}$$

*where, $\boldsymbol{\sigma}$ Sigmoid function, $i_t$ input gate, $f_t$ forgot gate, $g_t$ cell gate, $o_t$ output gate, $C_t$ cell state, $h_t$ hiden state, $x_t$ input at time $t$, $W$, $U$, learnable weights, $B$ bias, Hadamard or entry wise product, tanh hyperbolic tangent function.*

Anomaly detection is of great importance for video surveillance structures. Most of the structures that are proposed use techniques like Convolutional Neural Network (CNN) and LSTM (Long Short- Term Memory) networks to efficiently train the machine in an effort to locate the anomalies in each supervised in addition to unsupervised manners. The supervised learning approach emphasizes the use of the present understanding approximately a specific anomaly to teach a machine even as the unsupervised learning approach, on the other hand, attempts to learn normality instead of learning abnormality. This means that if a massive deflection is visible from ordinary behavior, it provokes abnormality. The convolutional LSTM basic Architecture is shown in figure-5.

## 2. Literature Review

In recent times, many deep convolution neural networks have been proposed [9, 13, 29-32] to extract high-level features by learning the temporal regularity in video clips. Deep cluster-based anomaly detection techniques are introduced in [33-35]. Although deep learning technology has made significant progress in several other fields, there are few reports on deep learning methods used to detect video anomalies. However, for deep learning-based video anomaly detection techniques, such as abnormal activity detection or abnormal object detection, there are few specialized investigations have been made and many researches are already cried out as shown in Table 1.

**Table 1:** *Summary of Recent work*

| *Year wise Work proposed* | *Broad area and Challenges* |
|---|---|
| *Wan et.al, 2021[37]* | *Anomaly Detection in Video Sequences: A Benchmark and Computational Model* |
| *Lu et.al, 2020[21]* | *Challenge is providing only Few-Shot Scene and detect anomalies from huge data.* |
| *Bansod et.al,2019 [39]* | *Transfer learning for video anomaly detection* |
| *Zhu et.al,2020[51]* | *Video Anomaly Detection for Smart Surveillance* |
| *Ramachandra et al.2020[53]* | *A Survey of Single-Scene Video Anomaly Detection* |
| *Basora et.al, 2019 [38]* | *Recent Advances in Anomaly Detection Methods Applied to Aviation* |
| *Chalapathy et.al, 2019 [40]* | *Deep-learning-based anomaly detection techniques for various domains* |
| *Chong et.al,2017[14]* | *Abnormal Event Detection in Videos using Spatiotemporal Autoencoder* |
| *Sultani et.al. 2018[17]* | *Real-world Anomaly Detection in Surveillance Videos.* |
| *Kiran et.al.2018[18]* | *An overview of deep learning-based methods for Unsupervised and semi-supervised anomaly detection in videos.* |

The team argue that the low recognition performance of these baselines reveals that the dataset is very challenging and opens more opportunities for future work. Bharathkumar Ramachandra et al, 2020 [53] reported on a survey of single-scene video anomaly detection. VIDEO anomaly detection is the task of localizing anomalies in space and/or time in a video. They have provided a comprehensive review of research in single-view video anomaly detection. The authors built an intuitive taxonomy and situated past research works in relation to each other. The pixel-level criterion of follows: Given the predicted anomaly score map St corresponding to the $tth$ frame of a test video, the frame is counted as a true positive frame. Sijie Zhu (2020) [56] noted that in modern intelligent video surveillance systems, automatic anomaly detection through computer vision analytics plays a pivotal role. Video anomaly detection has been studied for a long time, while this problem is far from being solved. Since real-world anomaly events happen with low probability, it is hard to capture all types of anomalies. This problem is typically formulated as unsupervised learning, where the models are trained with only normal video frames and validated with both normal and anomaly frames. UCF-Crime is currently the largest anomaly detection dataset with realistic anomalies, which contains thousands of anomalies and normal videos. The analysis involved 5 popular benchmark datasets.

Boyang Wan et al. [37] noted that Anomaly detection aims to distinguish abnormal and normal activities as well as anomaly categories in video sequences. Existing databases only provide video-level labels in training set, which makes it infeasible to learn anomaly detection models in a fully-supervised manner. By tracking all moving objects in a video sequence, the anomaly event is detected by considering different levels of spatiotemporal contexts. 100 video sequences are collected for each abnormal category, making it the largest database for anomaly detection to date. Anomaly detection attempts to automatically predict abnormal/normal events in a given video sequence. In the proposed multi-task deep neural network, the local spatiotemporal features are first extracted by an inflated 3D convolutional network from each video segment.

A group led by Zhuang-Zhuang Wang at the School of Computer Science and Engineering, [20] noted that the process of the proposed small-object detection algorithm is divided into four parts: input, backbone network, neck network, and head. The team continue to use PANet and the spatial pyramid pooling layer structure to fuse the feature information of feature maps of different sizes. It is believed that improving the accuracy of small-target detection by enhancing the resolution of the image will increase the number of calculations of the network. Small- object detection plays a key role in many tasks such as identifying traffic signs or pedestrians that are almost invisible in low-resolution images. This study introduces the FFT module to complete images SR and uses Darknet53 combined with dense block to extract small target features.

MyeongAh Cho et.al [54] reported in 'Unsupervised Video Anomaly Detection via Normalizing Flows with Implicit Latent Features' that surveillance anomaly detection finds abnormal events such as traffic monitoring, accidents, and crime using the petabytes of videos from CCTVs. They suggest distribution learning with normalizing flow models using static and dynamic features obtained from implicit two-path AE. They propose an ITAE that implicitly focuses on static and dynamic features. These two encoders generate a higher reconstruction error than one-path encoder for scenes with abnormal motion or appearance and perform anomaly detection better. Novel method for learning to detect anomalies in videos with only a few frames of video footage, which could have huge potential in real-world applications.

A research team led by Yiwei Lu of the University of Manitoba [21] have introduced a new problem called few-shot scene-adaptive anomaly detection. The group believe this new problem setup is closer to the real-world deployment of anomaly detection systems. Experimental results show that the proposed approach significantly outperforms other alternative methods. They consider the problem of anomaly detection in surveillance videos. Given a video, the goal is to identify frames where abnormal events happen. They learn a model that can quickly adapt to a new scene by using only a few frames from it. The researchers propose to learn few-shot scene-adaptive anomaly detection models. Target Methods K=1 K=5 K=10. Ped1 Fine-tuned 76.99 77.85 78.23 Ours 79.94 80.44 78.88 Ours 80.6 81.42 82.38. Ped2 Fine-tuned 85.64 89.66 91.11 Ours 90.73 91.5 91.11 Ours 91.19 91.8 92.8. Aspects of the findings appear to offer an alternative view to previous work in this subject: "In the train/test split used in, both training and test sets contain videos from the same set of 13 scenes.

# 4. Challenges in Anomaly Detection

At the abstract level, anomalies are defined as patterns that do not correspond to expected normal behavior. Therefore, an easy way to detect anomalies is to identify region that represent normal behavior and declare any observations in the data that are not related to that normal behavior that region is like an anomaly, but there are several factors that make this apparently simple method very difficult. Here is a list of the challenges we encountered in the research process 1) it is difficult to define a region that comprehends all probably normal behavior. It is due to the boundary between normal and outlying behavior is often not precise. Thus, an anomaly observation situated close to the boundary may be normal or abnormal. 2) While anomaly is arising due to the malicious deed, these malicious rivals often accommodate themselves to appear abnormal behavior as normal so the task of labeling normal behavior region becomes more difficult. 3) Current notion of normal behavior might not be sufficiently represented in the future. Means For different application domains the exact view of anomaly is dissimilar. For example, a small variation in normal reading might be diseases in the medical domain. Whereas, in the stock market domain small variation might be considered normal. This makes it complex to apply a particular domain technique to another domain. 4) The exact notion of an anomaly is different for different video surveillance video data-sets. 5) Availability of labeled data for training/validation. 7) Action pattern variations within the same class Environmental variations and noise Normal behavior keeps evolving. And the Actual Challenges found while performing experiments are: 1) Limited labeled data 2) Ambiguous definition of abnormal 3) Expensive feature engineering steps. 4) Abnormal events are challenging to obtain due to their rarity. Massive variety of abnormal events, manually detecting and labeling such events is a difficult task that requires much manpower. 5) Small objects in the wide data-set may behave like normal objects rather than anomalous.

# 5. Proposed Time Distributed ConvLSTM

The TDSTConvLSTM model work on context dependent and compare the result with benchmark data-set. The main objective of the research is to detection, localize and identify the anomaly object to provide security in surveillance video. Detecting an anomalous event in long sequence video is challenging. Due to the ambiguity of how strongly such events are defined, the accuracy of the object recognition in the video improves in order to achieve a better result with the reference data set. Achieve a better result with benchmark data-set. Optimize the model parameters and Compare results with benchmark data-set. Achieve better feature identification and representation. Train a model which is able to detect anomalies which are not significantly distinct from normal events.

### 1) Working of the Model:

A ConvLSTM is a variant of LSTM recurrent network. In ConvLSTM the internal matrix multiplication is replaced by convolutional operation, which allows the data through the ConvLSTM cells to keep the input dimension rather than just a one-dimensional vector with a function. The ConvLSTM also replaces fully connected layer operators with convolution operators [23-25, 41]. Figure-6 shows the final encoding and decoding for prediction network of Time distributed ConvLSTM. ConvLSTM use convolution operators for input to state and state to state connections. By replacing the convolution operators with an LSTM memory cell, the ConvLSTM model can therefore know which information from the previous state of the cell should be "remembered" or "forgotten" with the help of its forgetting gate. ConvLSTM also determines what information should be saved in the current state of the cell. The ConvLSTM process is described similarly to equations (1-5) is used to calculate the LSTM storage space. We have used these equations of basic Convolution LSTM and created new equations for hybrid Time Distributed ConvLSTM (eq1 -eq6).

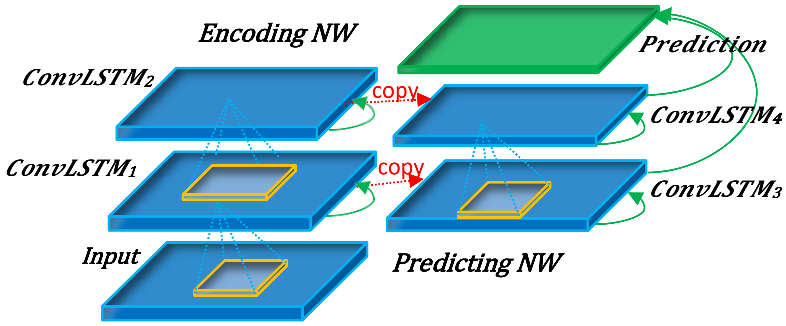



**Figure 6:** *Time Distributed ConvLSTM encoding and decoding for prediction*

The difference with TimeDistributed ConvLSTM is that the input vector $xt$ is supplied as images (i.2D or 3D matrices), with each weight in the connection being replaced by convolution filters. The intermediate state (of time steps) is analogous to the movement between frames.

## 2) Algorithm of working model

| **Algorithm 1** Training algorithm for Time Distributed ConvLSTM Encoder–Decoder |
|---|
| **Input:** Set of the sequential input range $X = \{\vec{X_1}, \vec{X_2}, \vec{X}, \ldots, \vec{X_t}\}$ |
| **Output:** Set of the sequential output range $Y = \{\vec{Y}, \vec{Y_2}, \vec{Y_3}, \ldots, \vec{Y_t}\}$ |
| Initialize network parameters by Xavier initializer $W \sim U\left[-\frac{\sqrt{6}}{\sqrt{X_t + X_{t+1}}}, \frac{\sqrt{6}}{\sqrt{X_t + X_{t+1}}}\right]$ |
| **while** the loss has not converged do |
| Compute loss between X and Y using Euclidean distance $E(X, Y, t) = \parallel I(X, Y, t) - fw(I(X, Y, t)) \parallel_2$ |
| Update parameters by ADAM optimizer |
| **end while** |

Here Xavier algorithmic used for weight initialization and Adam optimization used with a learning rate of 0.0001. When the learning loss stops decreasing, we attenuate it to 0.00001 and set the epsilon value to 0.000001. The distance between the frames are matured with the Euclidean distance algorithm. Then after reconstruction cost is calculated.

## 3) General Architecture:

The TimeDistributed layer provides exactly what we need. The generated allConv2D blocks are trained for the recognition we want, so our frames are processed to recognize things that are not simple object recognition, but things that "change" from frame to another frame. Figure-7 shows the

Time Distributed spatiotemporal ConvLSTM autoencoder. It is a neural network trained to reconstruct the input data. The TSConvLSTM autoencoder consists of two parts, Encoder and decoder. The ConvLSTM Encoder is able to learn a valid representation of the input($X$), called $g(\phi)$ encoding. The last layer of the encoder is called the bottleneck($z$) and contains the input representation $f(\phi)$. The ConvLSTM Decoder is use bottleneck coding to reconstruct the input data $R = f(g((\phi)) = X'R$.

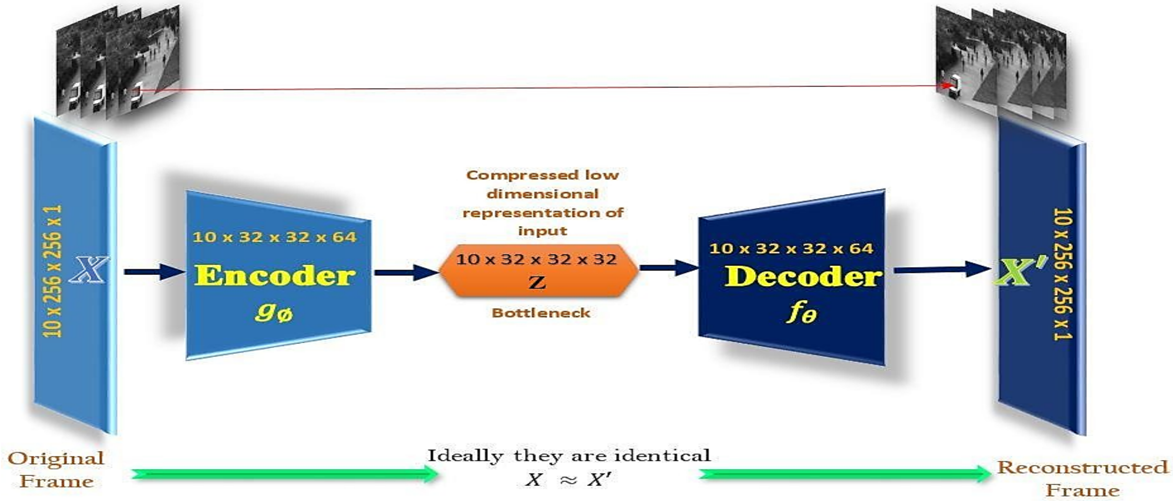



**Figure 7:** *Framework of Time Distributed spatiotemporal ConvLSTM N/W*

## 4) Collect Datasets & Analysis

It is expected that the trained anomaly detection model can be directly applied to multiple scenes with different perspectives; however, the existing data sets almost only contain videos recorded with fixed-angle cameras, which lack the diversity of scenes and angles. We summarize all anomaly detection data sets as follows: The data set Pedestrian 1 (Ped1) contains 34 training videos and 36 test videos, including 40 irregular events. All these unusual incidents involve vehicles such as bicycles and cars. Figures-8 and 9 show some examples of anomalies that are available in the UCSD data sets Ped1 and ped2.

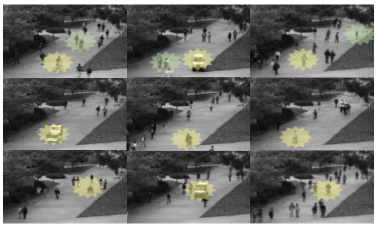


**Figure 8:** *UCSD PED1 dataset-Unlike normal training video streams, this scene consists of a small cart, Car, tempo, Cyclist, Wheelchair as well as a person skater*

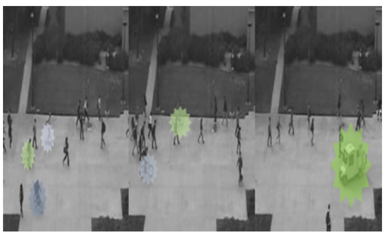


**Figure 9:** *UCSD PED2 dataset-Unlike normal training video streams, this scene consists of a small cart, Car, tempo, Cyclist, Wheelchair as well as a person skater*

**Figure 10:** *Shanghaitech dataset with different classes of anomalies*

The data set Pedestrian 2 (Ped2) contains 16 training videos and 12 test videos, including 12 abnormal events. The definition of Ped2 anomaly is the same as Ped1. The ShanghaiTech dataset is a collection of 13 scenes with complex lighting conditions and camera views in Shanghai University of Technology. It consists of 437 videos, each with an average of 726 frames. The training set contains 330 normal videos, and the test set contains 107 videos and 130 anomalies. Unusual events include unusual patterns on campus, such as cyclists or cars. https://svip-lab.github.io/dataset/campus_dataset. Figure-10 shows some of the available anomaly examples in the ShanghaiTech dataset.

**Table 2:** *Video Anomaly Detection Dataset*

| Dataset | Total Frames | Training Frames | Testing Frames | Anomalous Events | Anomaly Type | Irregularity | Regularity | Ground Truth | Resolution |
|---|---|---|---|---|---|---|---|---|---|
| UCSDPed1 | 14,000 | 6,800 | 7,200 | 54 | 5 | 4,005 | 9,995 | Spatial, Temporal | 238 x 158 |
| UCSDPed2 | 4,560 | 2,550 | 2,010 | 23 | 5 | 1,636 | 2,924 | Spatial, Temporal | 360 x 240 |
| CUHKAvenue | 30,652 | 15,328 | 15,324 | 47 | 5 | 3,820 | 26,832 | Spatial, Temporal | 640 x 360 |
| Shanghai-Tech | 3,17,398 | 2,74,515 | 42,883 | 130 | 13 | 17,090 | 3,00,308 | Spatial, Temporal | 856 x 480 |

CUHK Avenue data set contains 16 training videos and 21 test videos, a total of 47 anomalies Incidents, including throwing objects, loitering and running. The size of the person can vary according to the position and angle of the camera. The subway data collection takes a total of 2 hours. There are two categories: entry and exit. Unusual experiences include going in the wrong direction and wandering. More importantly, this data set was recorded indoors, and the above data was recorded outdoors. https://data.world/datasets/subway. The metadata of the benchmark datasets is shown in Table 2 and the distribution of the dataset by domains is shown in Figure-11.

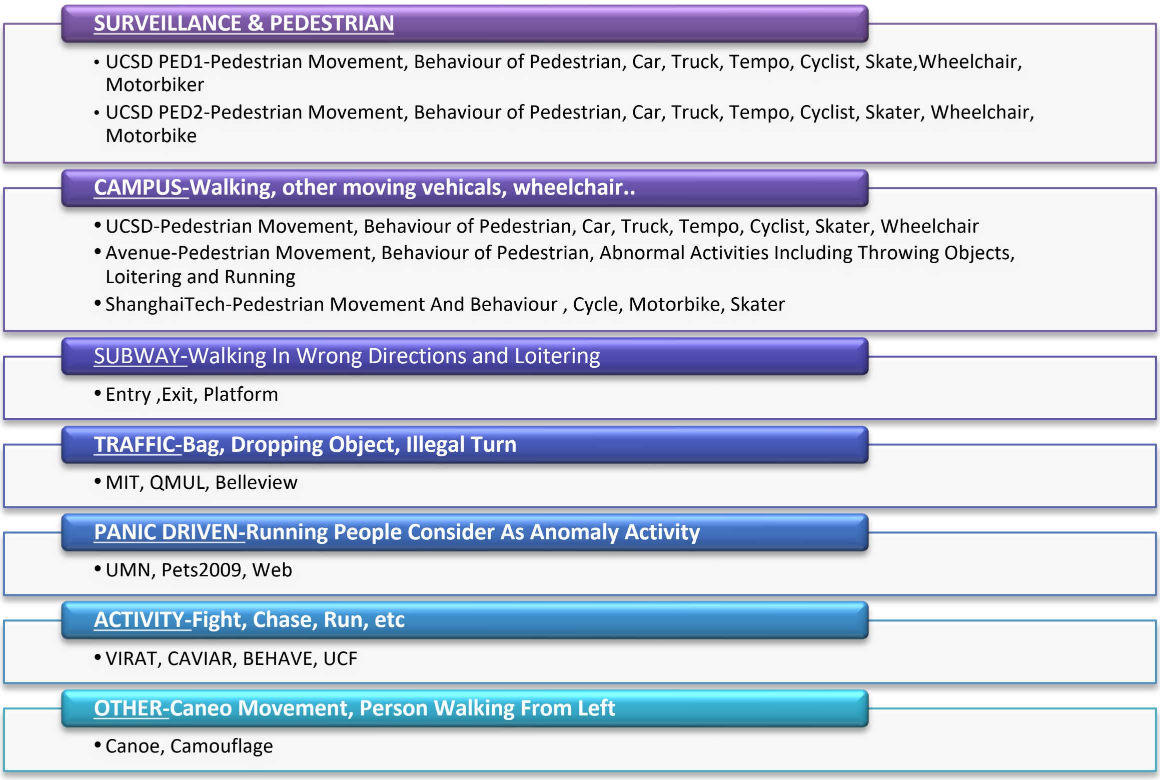



**Figure 11:** *Datasets available by domain type*

## 5) Hybrid TimeDistributed ConvLSTM:

The workflow of the existing approach (Figure-12) involves streams (spatial and temporal) that learn features during the encoding after which decode to generate reconstructed sequences from the video frame sequences. During Training our autoencoder trains on normal records through reconstruction so we are considering this as an unsupervised learning approach. When an abnormal event occurs, the corresponding reconstruction error score is higher than the normal data because the model did not follow the irregular pattern through training. Moreover, Features of the spatial model's convolution layer to identify pathways that could help better understand and represent the learning process of the model at the object level.

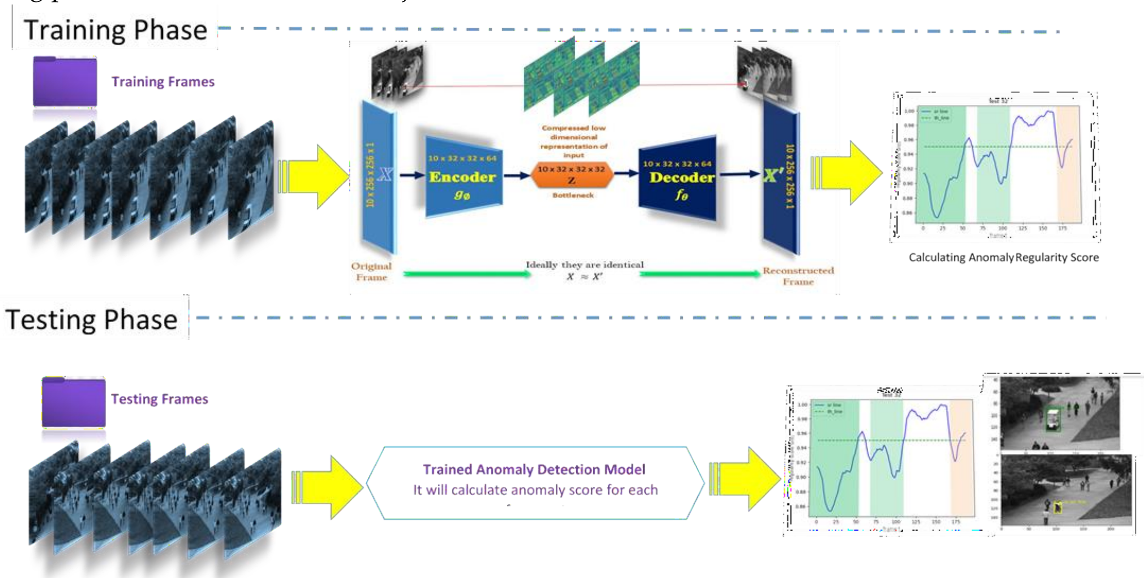



**Figure 12:** *Pipeline of the Time Distributed spatiotemporal ConvLSTM*

### 6) Reconstruction based on auto-encoder:

The input to the two-channel network is normal video frames. We trained the model and calculated the reconstruction error between the original frame sequence and the newly reconstructed frame sequence. The reconstruction error is used to calculate the regularity score, which can be further evaluated for the recognition performance. Our approach generates reconstruction errors for the spatial and temporal flows during the testing phase and then merges them accordingly. Our approach has three main steps.

**a) Pre-processing:** Several video clips were used to build and test the model, which vary in size, recording time and resolution. We decomposed the anomaly detection dataset into a sequence of video frames and normalized the video frame size to 256 x 256 pixels. To ensure that all input video frames are at the same scale, we compute the average pixels of the training image. We then subtract each frame from the global average for normalization. We also convert the image to grayscale to reduce the dimensionality. Due to the large number of training parameters and the limited training dataset, we used the data augmentation method [29] to expand the training dataset in time. frames (for example, in a cube at step 1, all T-frames are sequential, while at step 2 and step 3, the cubes skip one and two video frames, respectively) example of various skipping stride of s1[10]= [1,2,3,4,5,6,7,8,9,10] and s2[10]=[1,3,5,7,9,11,15,17,19,21], s3[10]=[2,4,6,8,10,12,14,16,18,20]. Once data augmentation done, we move for training and building model.

**b) Training and testing:** The encoder accept a sequence of input frames in chronological order. Two encoders are created here, a spatial encoder and a temporal encoder. The abstract features collected by Spatial Encoder are transmitted to Temporal Encoder to identify motion encoding.

Training and test datasets First, the data set is divided into two parts: train set and test set. The training set only contains videos with regular movements, and we included mixed videos with regular and irregular movements in the test set. The working principle of this model is as follows. Contains only normal motion patterns, no videos with abnormal motions,

$$X_{taining} \in \mathbb{R}^{N_{training} \times R \times C}$$

In given a frame testing sequence from video, which likely to contain anomalies,

$$Y_{testing} \in \mathbb{R}^{N_{testing} \times R \times C}$$

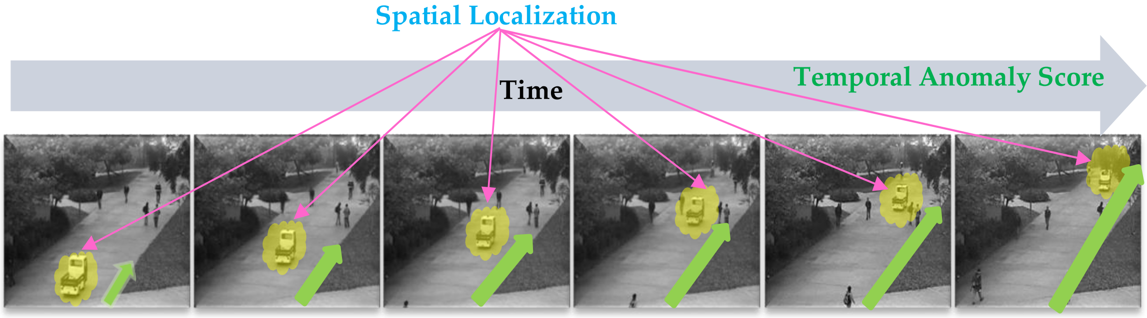



**Figure 13:** *Imagining anomalous regions and temporal anomaly score*

The problem is to assign an anomaly score to the temporal variation (time change of each frame), and a space score (spatial score) to locate anomalies in space, as shown in the below figure-13. When there is no direct information or guidance available for the positive rare category, the task of detecting anomalies is generally considered to be unsupervised; however, there are no abnormal samples. For,

$$R = \{X_i, Y_i\}, i \in [i, N], where\ Sample\ Y_i = 0$$

In below equations consider "⊛" for convolution operation, and "∘" for Hadamard product. LSTM is the special form of ConvLSTM, so ConvLSTM can be used as the LSTM. Input and state at a timestamp are 3D tensors. Convolution is used for both the input-to-state and state-to-state connection. Hadamard product is used to keep the constant error carousel property of cell.

$$i_t = \sigma(W_i \circledast x_t + U_i \circledast h_{t-1} + V_i \circledast C_{t-1} + b_i) \quad (eq1)$$
$$f_t = \sigma(W_f \circledast x_t + U_f \circledast h_{t-1} + V_f \circledast C_{t-1} + b_f) \quad (eq2)$$
$$g_t = tanh(W_g \circledast x_t + U_g \circledast h_{t-1} + b_g) \quad (eq3)$$
$$o_t = \sigma(W_o \circledast x_t + U_o \circledast h_{t-1} + V_o \circledast C_{t-1} + b_o) \quad (eq4)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ g_t\, x \quad (eq5)$$
$$h_t = o_t \tan(C_t) \quad (eq6)$$

As shown in above Figure-13 we have done experimentations on pedestrian dataset so we have used here UCSD Ped1 & Ped2, for Campus Test Avenue and ShanghaiTech are used in experiment. Once we collect the benchmark datasets prepare it for training. So, here preprocess is required to convert video data into frame sequence. So, the whole video is divided into sequence of frames of each size 10fps through Sliding window technique, for example The UCSD Ped1 have Video-1 we have converted into consecutive 200 frames of size 256 x 256 and same way we have generated all frames for all the datasets. Also, we have use scale the frames between pixel value 0 and 1 by dividing each pixel by 256.

**c) Learning Features:** we use spatial sequence to study the appearance, and we use temporal sequence to find out the temporal consistency in adjacent video frames. The temporal model consists of three parts:

convolutional layers, decoding layers, and the convolution long short-term memory. (ConvLSTM2D). A Convolutional Layer is used to examine the spatial or behavioral characteristics of each frame. The convolution layer is used to learn spatial features, and the Deconvolution layers used to restore the original input size and the ConvLSTM layer outperforms the temporal rules of video. Our spatial model is similar to the temporal model, but the spatial model lacks the Conv LSTM layer, and its input is represented as a single frame rather than sequential frames.

**i) Spatial model:** Table 3 demonstrations the detailed configuration of the proposed spatial model. It consists of only three layers of folding followed by two layers of deconvolution to improve efficiency. Since anomaly detection focuses more on low-level contours and edge features, only the spatial model is used and uses three layers of convolution for feature extraction. On the other hand, the role of the deconvolution layer is to generate reconstructed video frames and to condense the sparse inputs through operations with multiple filters. Therefore, the spatial size of the output feature maps of a deconvolution layer is larger than the spatial size of its corresponding inputs. Therefore, we extract the appearance of the person in the video through the three-layer convolution layer and restore the initial input dimensions through the two connected deconvolution layers using the layer parameter set through the training process. During the training phase, the learnable parameters were updated in the direction of minimizing the loss function. We use the loss of MSE based on the sigmoid function. By calculating the partial derivatives of the loss function, we were able to update the parameters in an Adam scheme. The process of learning feature is an important step in training a model. At the encoding stage, the model learns the spatial characteristics of the observed object in the video frame and important background information in the observed scene. The inputs and the spatial architecture is relatively simple. The algorithm for visualizing the feature map must make the "black box" of the spatial model transparent, understand the learning process of the model, and trust the final detection result.

**ii) Temporal model:** The temporal model can have the same formula, but different classes depending on the LSTM requirements. To better examine temporal consistency in adjacent frames, we added three Time Distributed ConvLSTM layers between convolution and deconvolution layers (Figure-14).

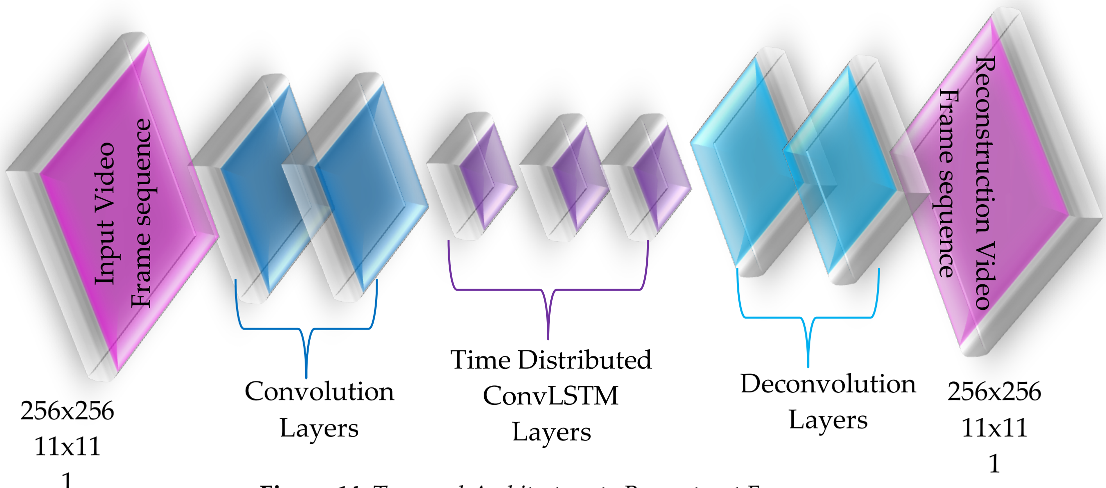



**Figure 14:** *Temporal Architecture to Reconstruct Frame sequence*

All three layers are equal, and the main difference lies in the number of convolution kernels. ConvLSTM is used to capture the spatiotemporal relationship in the data set. The difference between ConvLSTM and LSTM is that ConvLSTM changes the way LSTM switches from Hadamard product to convolution. The equations 1-5 are re-written as below equations (eq1-eq6). The proposed TSConvLSTM can predict the progress of a video sequence from a large number of input frames. Afterwards, the Regularity Score (RC) estimates come from a set of predicted reconstruction errors (RE). Abnormal video sequences produce lower regularity Scores because they deviate more from the actual sequence over time. The model uses composite structures and examines the influence of conditions on more meaningful learning. The best model is selected based on the accuracy of reconstruction and prediction. The 2DConvLSTM model performs qualitative and quantitative evaluation and displays the competitive results on the anomaly detection data set. The 2DconvLSTM blocks have proven to be an effective tool for modeling and predicting video sequences.

**d) Reconstruction & Regularity Score:** After obtaining the reconstructed sequence of the video frame, we calculated its reconstruction error between the original video frame sequences and the reconstructed frame sequences to model the probability distribution of the standard data. In our proposal, the reconstruction is a stochastic process which considers the distance between the reconstruction and the original video frame and the variability of the distribution itself. Here to find the distance between the two pixels of original and reconstructed frame Euclidian Distance is used. To qualitatively analyze whether our model can detect anomalies well, we used the regularity score table to indicate our model's ability to detect anomalies and the smoothness score corresponds to the normal level of each frame of the video sequences.

i) In run-through, we first counted the reconstruction error of the video frame before getting the regularity score $sr(t)$. We calculated the reconstruction error of the pixel intensity value $I$ at the location (x, y) in frame $t$ as follows:

$$E(x, y, t) = \| I(x, y, t) - fw\big(I(x, y, t)\big) \|_2 \qquad (eq7)$$

Where $f$ represents our two-stream model. We calculate the Euclidean distance between the initial pixel of the $t - th$ frame and the pixel of the reconstructed frame as the reconstruction error of the pixel. For each frame, we compute the reconstruction error probability by summing up all the pixel-based probabilities.

ii) Reconstruction error of consecutive frames of unlabeled frame sequence

$$E(t) = \sum_{(x,y)} E(x, y, t) \qquad (eq8)$$

iii) $Sequence\_Reconstruction\_Cost(t)$

$$SrC(t) = \sum_{t'=t}^{t+10} E(t') \qquad (eq9)$$

iv) Abnormality Score $Sa(t)$ is scaling between 0 and 1

$$S_a(t) = \frac{Sequence\ Reconstruction\ Cost(t) - Sequence\ Reconstruction\ Cost(t)_{min}}{Sequence\ Reconstruction\ Cost(t)_{max}}$$

$$S_a(t) = \frac{S_rC(t) - S_rC(t)_{min}}{S_rC(t)_{max}} \qquad (eq10)$$

v) Regularity Score $Sr(t)$-The abnormality score $Sa(t)$ corresponds to the level of abnormality of each frame in the video, which plays a role in indicating the confidence of detection results. On the other hand, the regularity score Sa(t) corresponds to the level of normality can be defined as follows:

$$S_r(t) = 1 - S_a(t) \qquad (eq11)$$

Assume that the regularity score of the current frame is relatively low. If there is no abnormal event in the video frame, the regularity score of the frame should be high. Also, in other-hand, if the regularity score is low then the possibility of abnormality in the video frame is also relatively high.

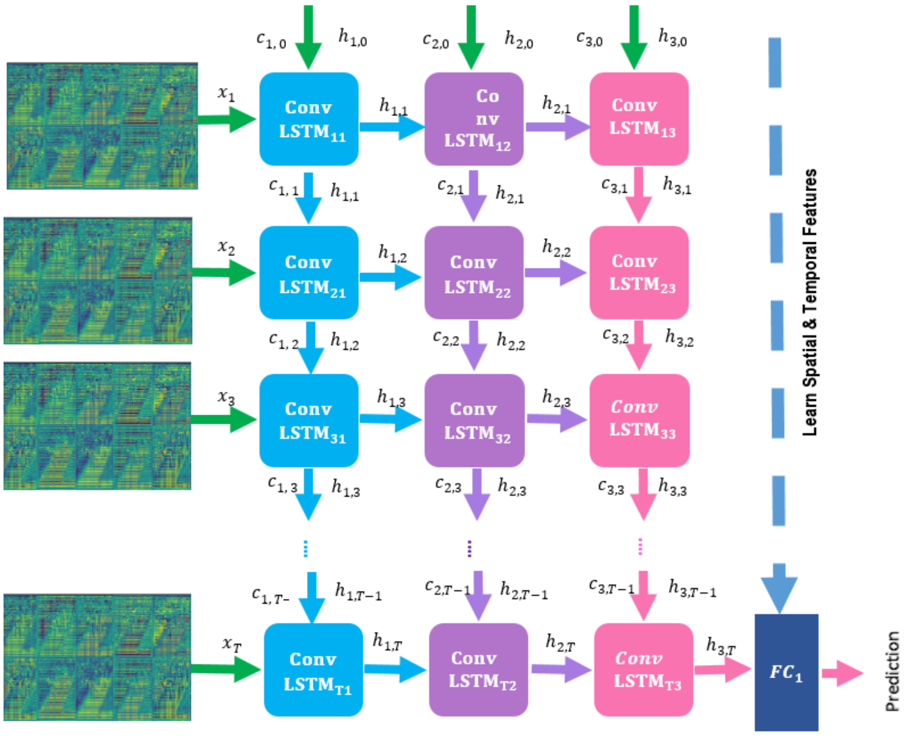



**Figure 15**: *Layer view of TSConvLSTM Model*

## 6. Experiment and Results:

### 1) Data Preprocessing and Model Configuration:

We analyze the overall performance of the proposed approach Time Distributed Spatiotemporal ConvLSTM mainly in relation to a surveillance dataset: UCSD Anomaly Detection Datasets, ShanghaiTech, and CUHK Avenue dataset. The UCSD anomaly detection dataset was collected with a fixed camera mounted on an elevated position overlooking the Campus pedestrian walkway. The density in the broad walkway can vary. Both the UCSD Ped1 and Ped2 datasets contain a train set and a test set. In particular, Ped1 contains 34 train video sequences and 36 test video sequences. The frame resolution is 256 × 256 pixels. In Ped1, people walk to and from the camera to create a foreshortening effect. UCSD Ped2 contains 16 train video sequences and 12 test video sequences with pedestrian traffic parallel to the camera plane. Frame resolution - 256 x 256 pixels. All frames in the train set are normal and contain only pedestrians. In addition to the normal frames, the test set had unusual frames in which cyclists, skaters, minivans, tempo, cars, wheelchairs, or people walking on the lawn were unusual. The second is ShanghaiTech Facility dataset to evaluate Time Distributed Spatiotemporal ConvLSTM method. The ShanghaiTech Campus dataset consists of 13 scenes with 107 items individual sets, each with complex lighting and camera angles. Each color block has a resolution of 856 x 480 pixels. As with the UCSD dataset, all train videos are normal and contain only pedestrians, and the test frames for each scene contain abnormal images All tests are performed on a dedicated server with a GPU with an Intel(R) Processor Core (TM) i7-7700HQ running at 2.80 GHz, 16 GB RAM, Nvidia GeForce GTX 1070 GPU running the Windows 10 operating system also with Ubuntu 16.04 it works well. We use the Python library, which is an open-source machine deep learning library for Python, to implement our anomaly detection architecture. We compared the s Time Distributed Spatiotemporal ConvLSTM with several advanced video anomaly detection baselines trained with regular videos only and focused on the default unsupervised learning setup. Anomaly detection in all datasets using the AUROC scale is presented in Table 3. Table 3 shows a comparison with other methods using four benchmarks. Our approach delivers competitive or superior results without using a pre-trained network, and Time Distributed Spatiotemporal ConvLSTM improves performance over the well-designed Bansod et.al, 2019 [39], ConvLSTM [14], ITAE [31], ConvLSTMAE [15]. In the CUHK database ConvLSTM [45], MLEPFP [46], ConvLatentAE [44], Wan et.al, 2021 [37] a prediction-based method that stores and updates normal query functions for each memory module.

### 2) Parameter Selection

**Table 3:** *Parameter Selection*

| **Parameter Selection:** Input Length, output Length, kernel Size, Type of Normalization, Output Nonlinearity |
|---|
| Conv kernel -( 11 x 11 x 1 x 128) |
| Conv kernel -( 5 x 5 x 128 x 64) |
| Conv kernel -( 3 x 3 x 64 x 32) |
| ConvLSTM-(3 x 3 x 32 x 256)<br>**Recurrent kernel** -(3 x 3 x 64 x 256)-TanH |
| ConvLSTM-(3 x 3 x 64 x 256) **– Recurrent kernel** -(3 x 3 x 64 x 256)-TanH |
| ConvLSTM-(3 x 3 x 64 x 256) **–Recurrent kernel** -(3 x 3 x 64 x 256)-TanH |
| Conv Transpose kernel -( 3 x 3 x 32 x 64) |
| Conv Transpose kernel -( 5 x 5 x 64 x 32) |
| Conv Transpose kernel -( 5 x 5 x 128 x 64) |
| Conv Transpose kernel -( 11 x 11 x 128 x 64) |
| Conv kernel -( 11 x 11 x 128 x 1)-Sigmoid |

Model Parameters the input videos are converted into sequence of frames and that individual frame converted to grayscale and resized to 256 x 256 pixels. A preliminary Conv-LSTM Encoder-Decoder baseline model was evaluated for use as reference in parameter selection. The baseline model utilizes an input and output length of five, and divides the image into non-overlapping patches. Using the model from [45,48] as reference, but with modification in filter size of 11 x 11, 5x5 and 3 x 3 and three Time Distributed ConvLSTM layers are used, while the total number of filters are 128, 64, 32 respectively to the encoder and transpose decoder to accommodate the larger frame sequence. The Time Distributed ConvLSTM units use recurrent activation sigmoid nonlinearities for the input, output and forget states, and tanH for the hidden and cell states. A sigmoid non-linearity is applied to the final Time Distributed convolutional layer. Due to case insensitivity the "same padding" method means, zeros evenly to the left/right or up/down of the input such that reconstruct frames has the same height/width dimension as the input frames is applied during all convolution operations to retain the frame size.

The baseline model is simpler than the complex Time Distributed ConvLSTM and utilizes only a future decoder. The parameters tested in variations of the TDSTConv LSTM model include the length of both the input and output timestamps, the filter size, and the final output non-linearity function, as shown in Table-1. A filter size of 3x3 was considered for capturing smaller motions, but was not as effective. The commonly used sigmoid nonlinearity function was tested at the final output, but ultimately, the parameters used by the baseline model were found to be the most effective. The parameters were applied to the proposed composite models and evaluated for accuracy with respect to the baseline model, as shown in Table 3. The composite models have a lower MSE per frame with the unconditioned model performing slightly better.

### 3) Evaluation parameter of Anomaly Detection Experiments

Quantitative Analysis: Frame-Level AUC. To better compare with other methods, all the experiments are carried out on the same work station with Intel Intel(R) Core (TM) i7-7700HQ CPU @ 2.80GHz 2.80 GHz, NVIDIA GTX 1070, and 16G RAM. If a frame contains at least one abnormal event, it is considered as a correct detection., Detection is compared to the frame-level ground-truth label. The area under the curve (AUC) is the evaluation metrics. Furthermore, some contemporary documents [9, 10] believe that the EER evaluation criteria are a severe sample imbalance between normal and abnormal events. Using EER as an indicator will be misleading in practical applications. We agree with this view and use AUC for evaluation, assuming that the local minimum within 50 frames belongs to the same abnormal event. A temporal window of 50 frames before and after distinct local minima is used to propose anomalous regions, as most anomalous activities are at least one hundred frames long. The proposed local minimum regions within fifty frames of each other are connected to obtain the final abnormal temporal regions. These minima are then considered to be a part of same abnormal event. We consider a detected abnormal region as a correct detection if it has at least fifty percent overlap with the ground truth/ fact table. According to Kozlov et al., 2013 A parameter-sweep at intervals of .05 is performed to determine the threshold parameter for the Persistence1D algorithm.

### 4) Efficiency Analysis of Model

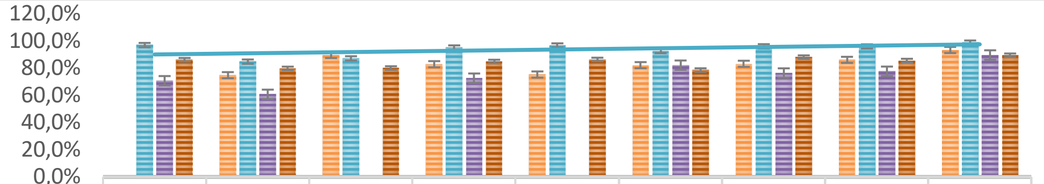
Table 3 presents the AUC of our method and a series of state-of-the-art methods [5, 6, 8, 11] on the Avenue, the UCSD Ped2, and the Subway Entrance and Exit datasets. As expected, our model performs the best performance on the avenue and subway entrance and exit datasets. In addition, although the version in Avenue and Ped datasets appears to be slightly lower than that in the other complicated architectures, it is still significantly higher than that of lightweight models and that single-level models. These results indicate that a multilevel model [8] or 3D indicator [6] can perform better in crowd scene, such as the UCSD Ped2 dataset. However, the time cost of these methods was also higher. Besides, comparing our spatial model and temporal model and the fusion model, temporal and spatial model have their advantages and disadvantages. Still, the fusion model performs better than the former two on all data sets.

### 5) Optimization and Initialization

The cost function of equation (eq9) was optimized with Adam optimization. Adam optimization is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments and it was empirically chosen as the most effective. After all the experiments the learning rate of $1e-4$, decay rate of $1e-5$ and epsilon $1e-6$ were finalized. The detailed parameters are shown in table 4. We used a mini-batch of five video sequences and trained the models for up to 25,000 iterations. Early stopping was performed based on the validation loss if necessary. The weights were initialized using the Xavier Weight Initialization. It automatically scales the initialization based on the number of input and output neurons to prevent the weights from starting out too small or large, and vanish or explode in magnitude. The input to-hidden and hidden-to-hidden convolutional filters in the TDSTConvLSTM units all use the same filter size.

**Table 4:** *The Final Parameter value settings of TDSTConvLSTM*

| Parameter Type and value | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Height | Width | Batch size | Learning Rate | Epoch | Optimizer | Stride | Loss | decay |
| 256 | 256 | 4 | $1e-4$ | 200 | Adam | 4 | MSE | $1e-5$ |




| | Doshi et.al, 2020[52] 2010 | Conv2d-AE[19] 2016 | Chong et.al,2017[14] 2017 | FFP+MC[55] 2018 | DeepOC[47] 2019 | Lu et.al, 2020[21] 2020 | ROADMAP[36] 2021 | MAML[21] 2021 | TDSTConvLSTM [Our] 2021-22 |
|---|---|---|---|---|---|---|---|---|---|
| UCSD PED1 | | 75,0% | 89,9% | 83,1% | 75,5% | 82,4% | 83,4% | 86,3% | 93,6% |
| UCSD PED2 | 97,2% | 85,0% | 87,4% | 95,4% | 96,9% | 92,8% | 96,3% | 96,2% | 99,0% |
| ShanghaiTech | 70,9% | 60,9% | | 72,8% | | 82,4% | 76,6% | 77,9% | 89,9% |
| Avenue | 86,4% | 80,0% | 80,3% | 85,1% | 86,6% | 78,8% | 88,3% | 85,8% | 89,7% |

**Figure 16 :** *Performance comparison AUC (%) of the anomaly detection result of TDSTCONVLSTM models with state-of-the-art methods on four benchmark datasets*

As shown in Figure-16 the area under the curve AUC of ROC (Receiver Operating Characteristic) curve is extensively used as a measure of the temporal localization of anomaly events. Since the anomaly detection can be taken into consideration as a binary type for each frame, the ROC curve is generated by applying distinct thresholds for the anomaly rating of every frame and calculating the TPR (True Positive Rate) and FPR (False Positive Rate).

## 6) Proposed Evaluation Criterion

As demonstrated by the experimental results on several video anomaly detection benchmark data sets and we achieve comparable performance contrast with the state-of-the-art unsupervised method with much less running time, indicating the effectiveness, efficiency, and robustness of our proposed approach. According to previous work [10, 12, 13] we evaluated our method by the area under the ROC curve (AUC). The ROC curve is obtained by varying the threshold value of the abnormality evaluation. A higher AUC value. Represents a more precise result of the anomaly detection. To ensure comparability between different methods, we calculated the AUC for the prediction at frame level. The Receiver Operating Characteristic (ROC) of the TimeDistributed Spatiotemporal ConvLSTM with UCSD

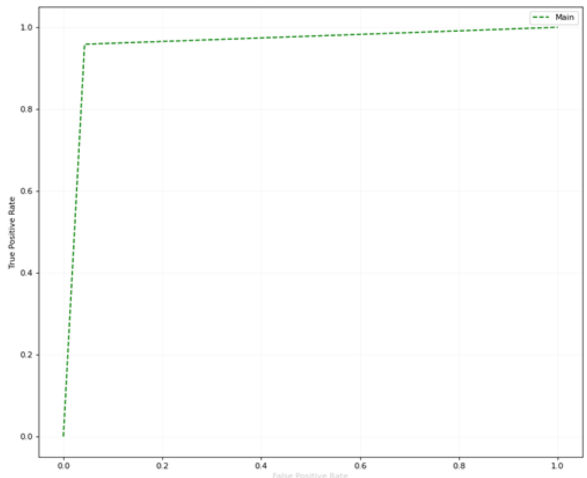



**Figure 17:** *ROC curve of a TDSTConvLSTM with UCSD Ped1 dataset*

Ped1 and Ped2, Avenue and ShanghaiTech dataset is shown in Figure-17. Figure-18 and 19 shows the Anomaly Regularity Score of Sets.

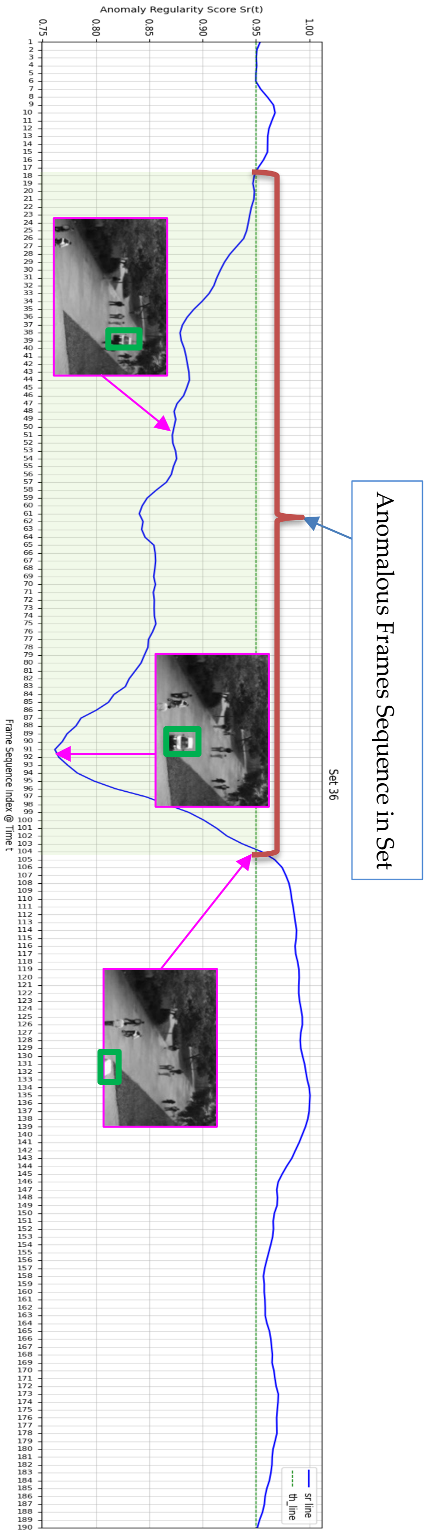



**Figure 18:** *Performance Analysis of set*

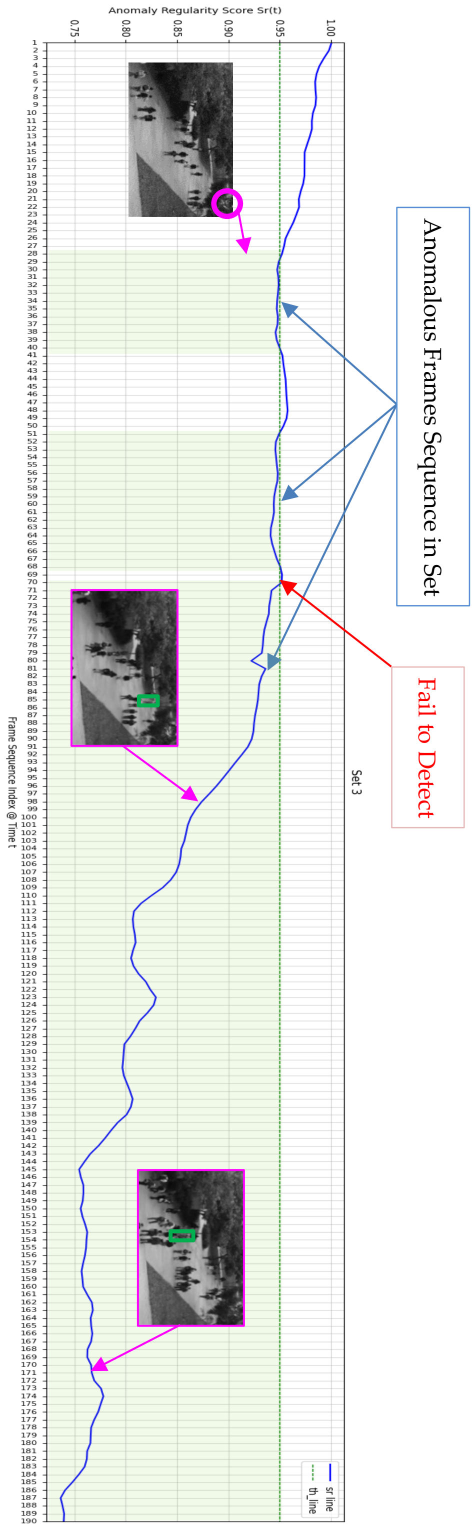



**Figure 19:** *Performance Analysis of set*

# Conclusion

In this paper, we proposed a TDSTConvLSTM modeling of normal features based in a supervised manner for video anomaly detection. We designed the TDSTConvLSTM to implicitly capture representative static and dynamic information of normal scenes without using a pre-trained network. For the complex normality, using the latent features of TDSTConvLSTM, through an experiment on standard benchmarks, TDSTConvLSTM demonstrated high effectiveness in scenes where motion is abnormal by learning dynamic information of normal scenes. Furthermore, the normality modeling of the TDSTConvLSTM feature achieved superior results, especially when the database is extensive and composed of diverse scenes. The proposed method can be expected to model a general distribution and solve practical problems through a vast number of real-world videos with Semi supervised learning. Our results suggest that the proposed method enables fast and reliable detection of abnormal events, with label-free identification of abnormal events. Using only the spatial or temporal stream cannot cause the best result. However, with the information from the two-stream fused, the model has improved efficiency compared with a single stream, while the accuracy is also competitive. It should also be noted that our current model is lightweight and does not consider the complete appearance and motion of the video scenario. Therefore, the training process in our method does not reconstruct all the changes in the properties of appearance and motion, and it may be weak compared with other techniques in a particular dataset for example, GMFC-VAE in Ped2.

# References

[1] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM computing surveys (CSUR). 2009 Jul 30;41(3):1-58.

[2] Patel P, Thakkar A. “Recent Advancement in Anomaly Detection in Surveillance Videos.” International Journal of Recent Technology and Engineering 8, no. 2 (n.d.): 964–70. doi:10.35940/IJRTE.B1759.078219.

[3] Feldstein S. The global expansion of AI surveillance. Washington, DC: Carnegie Endowment for International Peace; 2019 Sep 17.

[4] Chong YS, Tay YH. Abnormal event detection in videos using spatiotemporal autoencoder. InInternational symposium on neural networks 2017 Jun 21 (pp. 189-196). Springer, Cham.

[5] Chong YS, Tay YH. Modeling representation of videos for anomaly detection using deep learning: A review. arXiv preprint arXiv:1505.00523. 2015 May 4.

[6] B. Antic and B. Ommer, “Video parsing for abnormality detection,” in Proceedings of the International Conference on Computer Vision, Barcelona, Spain, November 2011.

[7] H. Mobahi, R. Collobert, and J. Weston, “Deep learning from temporal coherence in video,” in Proceedings of the 26 th International Conference on Machine Learning, Montreal, Canada, 2009.

[8] W. Li, V. Mahadevan, V. NJIToPA, and M. Intelligence, “Anomaly detection and localization in crowded scenes,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, pp. 18–32, 2014.

[9] K. Cheng, Y. Chen, and W. Fang, “Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression,” in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015.

[10] R. Mehran, A. Oyama, and M. Shah, “Abnormal crowd behavior detection using social force model,” in Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, June 2009.

[11] T. M. Hospedales, S. Gong, and T. Xiang, “A Markov clustering topic model for mining behaviour in video,” in Proceedings of the International Conference on Computer Vision, Kyoto, Japan, October 2009.

[12] Y. Cong, J. Yuan, and J. Liu, “Sparse reconstruction cost for abnormal event detection,” in Proceedings

of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 2011.

[13] W. Liu, W. Luo, "Future frame prediction for anomaly detection—a new baseline," 2018

[14] Y. S. Chong "Abnormal event detection in videos using spatiotemporal autoencoder," 2017,

[15] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional LSTM for anomaly detection," in Proceedings of the International Conference on Multimedia and Expo, Hong Kong, China, July 2017.

[16] K. P. Adhiya, S. R. Kolhe, and S. S. Patil, "Tracking and identification of suspicious and abnormal behaviors using supervised machine learning technique," in Proceedings of the International Conference on Advances in Computing, Communication and Control, Mumbai India, January 2009.

[17] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," 2018.

[18] BR Kiran, DM Thomas, R. Parakkal An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging.4(2):36, 2018.

[19] M Hasan, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS. Learning temporal regularity in video sequences. InProceedings of the IEEE conference on computer vision and pattern recognition pp. 733-742, 2016.

[20] Wang ZZ, Xie K, Zhang XY, Chen HQ, Wen C, He JB. Small-Object Detection Based on YOLO and Dense Block via Image Super-Resolution. IEEE Access.9:56416-29, Apr 9, 2021

[21] Lu Y, Yu F, Reddy MK, Wang Y. Few-shot scene-adaptive anomaly detection. InEuropean Conference on Computer Vision 2020 Aug 23 (pp. 125-141). Springer, Cham.

[22] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," arXiv preprint arXiv:1511.05440, 2015

[23] Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R. Learning from simulated and unsupervised images through adversarial training. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 2107-2116).

[24] S. Xingjian, Z. Chen, H. Wang, and D. Yeung, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in Advances in Neural Information Processing Systems, pp. 802–810.

[25] Essien A, Giannetti C. A deep learning model for smart manufacturing using convolutional LSTM neural network autoencoders. IEEE Transactions on Industrial Informatics. 2020 Jan 23; 16(9):6069-78.

[26] Chang, Y., Tu, Z., Luo, B., Qin, Q.: Learning spatiotemporal representation based on 3D autoencoder for anomaly detection. In: Cree, M., Huang, F., Yuan, J., Yan, W.Q. (eds.) ACPR 2019. CCIS, vol. 1180, pp. 187–195. Springer, Singapore (2020).

[27] Xu, D., Yan, Y., Ricci, E., Sebe, N.: Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput. Vis. Image Underst. 156, 117–127 (2017)

[28] Yan, M., Meng, J., Zhou, C., Tu, Z., Tan, Y.P., Yuan, J.: Detecting spatiotemporal irregularities in videos via a 3D convolutional autoencoder. J. Vis. Commun. Image Represent. 67, 102747 (2020)

[29] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)

[30] Poultney, C., Chopra, S., Cun, Y.L., et al.: Efficient learning of sparse representations with an energy-based model. In: Advances in Neural Information Processing Systems, pp. 1137–1144 (2007)

[31] Rifai, S., Vincent, P., Muller, X., Glorot, X., Bengio, Y.: Contractive autoencoders: explicit invariance during feature extraction. In: International Conference on Machine Learning (ICML), pp. 833–840 (2011)

[32] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: International Conference on Machine Learning (ICML), pp. 1096–1103 (2008)

[33] Blanchard, G., Lee, G.: Semi-supervised novelty detection. J. Mach. Learn. Res. 11, 2973–3009 (2010)

[34] Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning, pp. 478–487 (2016)

[35] A. Sodemann, M. P. Ross, B. J. Borghetti, A review of anomaly detection in automated surveillance, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (6)1257–1272.

[36] Wang X, Che Z, Jiang B, Xiao N, Yang K, Tang J, Ye J, Wang J, Qi Q. Robust Unsupervised Video Anomaly Detection by Multipath Frame Prediction. IEEE Transactions on Neural Networks and Learning Systems. 2021 Jun 4.

[37] Wan B, Jiang W, Fang Y, Luo Z, Ding G. Anomaly detection in video sequences: A benchmark and computational model. arXiv preprint arXiv:2106.08570. 2021 Jun 16.

[38] Basora L, Olive X, Dubot T. Recent advances in anomaly detection methods applied to aviation. Aerospace. 2019 Nov;6(11):117.

[39] Bansod S, Nandedkar A. Transfer learning for video anomaly detection. Journal of Intelligent & Fuzzy Systems. 2019 Jan 1;36(3):1967-75.

[40] R. Chalapathy, S. Chawla, Deep learning for anomaly detection: A survey, arXiv preprint arXiv:1901.03407

[41] Priyanka P. Patel, Dr. Amit R. Thakkar, "Understand Long Short Term Memory for Sequential Data", IJAST, vol. 29, no. 08, pp. 2482 - 2491, Jun. 2020.

[42] X. Chen, B. Li, J. Wang, Y. Zhao and Y. Xiong, "Integrating EMD with Multivariate LSTM for Time Series QoS Prediction," 2020 IEEE International Conference on Web Services (ICWS), 2020, pp. 58-65, doi: 10.1109/ICWS49710.2020.00015.

[43] Patel, Priyanka P., and Amit R. Thakkar. "A Journey From Neural Networks to Deep Networks: Comprehensive Understanding for Deep Learning." Neural Networks for Natural Language Processing. IGI Global, 2020. 31-62.

[44] Ionescu, Radu Tudor, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7842-7851. 2019.

[45] Medel, Jefferson Ryan, and Andreas Savakis. "Anomaly detection in video using predictive convolutional long short-term memory networks." arXiv preprint arXiv:1612.00390 (2016).

[46] Liu, Wen, Weixin Luo, Zhengxin Li, Peilin Zhao, and Shenghua Gao. "Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies." In IJCAI, pp. 3023-3030. 2019.

[47] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," IEEE transactions on neural networks and learning systems, 2019.

[48] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, W. Woo, Wangchun. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in NIPS, pp. 802-810, 2015.

[49] Patel, Priyanka, and Amit Thakkar. "The upsurge of deep learning for computer vision applications." International Journal of Electrical and Computer Engineering 10.1 (2020): 538.

[50] Datta, Leonid. "A survey on activation functions and their relation with xavier and he normal initialization." arXiv preprint arXiv:2004.06632 (2020).

[51] Zhu, Sijie, Chen Chen, and Waqas Sultani. "Video anomaly detection for smart surveillance." arXiv preprint arXiv:2004.00222 (2020).

[52] Doshi, Keval, and Yasin Yilmaz. "Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate." Pattern Recognition 114 (2021): 107865.

[53] Ramachandra, Bharathkumar, Michael Jones, and Ranga Raju Vatsavai. "A survey of single-scene video anomaly detection." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[54] Cho M, Kim T, Kim IJ, Lee S. Unsupervised Video Anomaly Detection via Normalizing Flows with Implicit Latent Features. arXiv preprint arXiv:2010.07524. 2020 Oct 15.

[55] Liu, Wen, Weixin Luo, Dongze Lian, and Shenghua Gao. "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.

[56] Zhu, S., Chen, C., & Sultani, W. (2020). Video anomaly detection for smart surveillance.