جامعة حلوان
HELWAN UNIVERSITY

كلية الحاسبات والذكاء الإصطناعي
Faculty of Computers & Artificial Intelligence

Early Anomaly Prediction in Surveillance Cameras for Security Applications

Graduation Project Documentation
2020/2021

Mario Emad, Michael Ishack,
Mohamed Ahmed, Mohamed Osama,
Mohamed Salah

Supervisor: Assoc.Prof. Ghada Khoriba

Computer Science Department
Faculty of Computers and Artificial intelligence

كلية الحاسبات والذكاء الإصطناعي
Faculty of Computers & Artificial Intelligence

جامعة حلوان
Helwan University

HELWAN UNIVERSITY
Faculty of Computers and Artificial Intelligence
Computer Science Department

Early Anomaly Prediction in Surveillance Cameras for Security Applications

A graduation project dissertation by:

[Mario Emad Kitshener (20170396)]

[Michael Ishack Georgy (20170399)]

[Mohamed Ahmed Mohamed Ali (20170412)]

[Mohamed Osama Nabawy (20170414)]

[Mohamed Salah Abdelkader (20170446)]

Submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Computers & Artificial Intelligence, at the Computer Science Department, the Faculty of Computers & Artificial Intelligence, Helwan University

Supervised by:

Dr. Ghada Ahmed

July 2021

جامعة حلوان
كلية الحاسبات والذكاء الإصطناعي
قسم علوم الحاسب

بناء نظام امني لتوقع الحوادث الاجرامية قبل حدوثها باستخدام الكاميرات الخارجية و الداخلية

رسالة مشروع تخرج مقدمة من:

[ماريو عماد كتشنر (20170396)]

[مايكل اسحاق جورجي (20170399)]

[محمد احمد محمد علي (20170412)]

[محمد أسامة نبوي (20170414)]

[محمد صلاح عبدالقادر (20170446)]

رسالة مقدمة ضمن متطلبات الحصول على درجة البكالوريوس في الحاسبات والذكاء الإصطناعي،
بقسم علوم الحاسب، كلية الحاسبات والذكاء الإصطناعي، جامعة حلوان

تحت إشراف:

ا.م.د. غادة احمد

يوليو / تموز 2021

Abstract

In the last decade, the number of surveillance cameras has increased significantly, with much research conducted to automate the process of surveillance, as humans cannot manage to monitor all these cameras individually, which may cause errors in public safety or abnormal situations. Also, humans may overlook key details in such abnormal behaviours in surveillance cameras. The proposed approach predicts abnormal behaviour using generative adversarial networks (GANs). GANs are trained using different datasets that contain various behaviours to predict future frames. These future frames are transmitted to a deep learning neural network to classify them as normal or abnormal activities, and future anomalies can be detected before they happen. Our initial results show that depending on the future frames extracted by the GAN model is possible, as these extracted frames either improve the accuracy of the detection model or do not affect it, but they can also be further enhanced to detect more frames at a longer duration and predict anomalies before they happen. Anomalies in surveillance will not only be detected but also predicted before they happen, which will result in the prevention of crimes, reductions in surveillance costs and a safer environment overall. Our results show that we can depend on the future frame predictions in the anomaly prediction model as the accuracy either have risen or did not change, the results have produced a 8.471 percent improvement on the Avenue Dataset and a 0.01 percent improvement on the Shanghai Tech Dataset , which then drive future works towards producing a higher quality frames and a higher number of frames in order to predict the anomaly way earlier.

Acknowledgements

First, We would love to thank our colleagues and fellow faculty members who have lend us any help in any form during our precious learning process. We would love to thank our supervisor Dr. Ghada Ahmed Khoriba for her amazing work, extended support for us, guidance and motivation. This project wouldn't have reached such a point without her. In addition, we would like to thank Eng. Abdallah Essam for his technical assistance in the Alexandria Library High Performance Computing Facilities as he had spent a lot of time helping us through the installation process and deployment. We would like to thank Mr. Yong-Hoon Kwon for his quick response and his written results to confirm that the model was working up to standards. We would like to also thank Alexandria Library HPC for providing us with their HPC Facilities. We would like to thank Academy of Scientific Research and Technology for accepting our graduation project into their Bedayaty program.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Surveillance Cameras

The development of surveillance camera systems has failed to deliver the promised deterrent effects or investigation case evidence, and their use has been unimpressive. Computer vision-enhanced camera networks that can deliver automatic real-time video analysis could be a feasible solution to practical camera monitor needs.
Surveillance has evolved from a human-based activity to one dominated by camera technology, thanks to the availability of less expensive cameras. Despite long-standing concerns about the consequences of video monitoring on society, cameras continue to be a popular choice for addressing a variety of security problems, and cameras are now found in a variety of public and private locations (Adams and Ferryman; 2015; La Vigne et al.; 2011; Sandhu; 2019; Scheitle and Halligan; 2018; Surette; 2014)
Human-monitored surveillance cameras have been proven to be effective in some contexts for some crimes, according to evaluations of police surveillance cameras, but no consistent beneficial results have been observed. According to a recent meta-analysis (Piza et al.; 2019), public space camera networks are linked to a slight but considerable reduction in crime, with the largest and most consistent reduction being observed in car parks.
Most of the existing research on surveillance cameras, on the other hand, has focused on their proactive value in crime prevention and reduction. Reactive applications have received relatively little research (Ashby; 2017). Beyond crime prevention, the use of surveillance camera footage for investigations is a surveillance camera application that has been described as having a lot of potential but has not been thoroughly investigated.
Beyond crime prevention, the use of surveillance camera footage for investigations is a surveillance camera application that has been described as having a high potential but has not been thoroughly investigated. Due to a long belief that surveillance camera networks are better for investigations than for crime reduction, little research has been done on the topic.

1.2 Problem of Surveillance Systems

Surveillance camera systems, according to (Honovich; 2019) , are better suited for crime solving rather than crime reduction and should be employed in crimes when offenders undertake pre-crime risk assessments.
(Honovich; 2019) stated that camera network operators should focus on

solving rather than discouraging crime, citing the widespread usage of surveillance cameras in the private sector, which are justified as crime investigation instruments. Even through this early warning, current studies focused on understanding crime prevention effects, and research on the use of surveillance cameras for investigative purposes remained limited, regardless of the fact that police routinely request surveillance video when conducting investigations when it is available (Morgan and Coughlan; 2018) .

Previous research on the value of surveillance cameras for investigations is frequently not robust. A significant portion is based on journalistic research and is published as news pieces (Ashby; 2017; Davenport; 2007; Bulwa; 2007; Edwards; 2008)

Surveillance cameras may be beneficial in criminal investigations because they can immediately assist in answering two fundamental investigative questions: what happened and who was involved. Video recording can allow investigators to observe a whole occurrence and verify or dispute other evidence or testimony, as well as provide reviewable proof about what happened and who did it. It is not necessary for footage to result in an arrest for it to be valuable, and the removal of a suspect or a crime is also socially beneficial.

In what situations are surveillance cameras most likely to come in handy? Police perceptions of the utility of surveillance camera footage have been divided in the past. Some police officers have described it as extremely beneficial, while others have described it as counterproductive—to the point where (Ashby; 2017) writes that some have advocated for the elimination of human monitoring. However, surveillance camera video appears to be effective in enhancing detection for a variety of crimes, including robbery and violent crimes (Ashby 2017). According to an Australian survey of police investigators, 9 out of 10 officers place a high importance on camera footage Dowling et al. (2019). When footage was provided, three out of four polled investigators thought it was useful or very useful. They also thought camera footage was particularly valuable in the early phases of investigations and preferred it for investigating assaults, while admitting that surveillance video, in their opinion, raised clearance rates for theft, burglary, and property damage the most (Dowling et al.; 2019). Identifying subjects was the most common purpose indicated, followed by creating leads, corroborating witness and suspect statements, and assessing whether or not a crime had happened. Overall, the evidence demonstrates that surveillance cameras can be effective investigation tools for a wide range of crimes.

1.3 Human Monitoring and its downsides

The number of cameras in many networks outnumbers human capacity to efficiently monitor them, which is one reason for the gap (Hesse; 2002; Prenzler and Wilson; 2019; Welsh et al.; 2015). In practicality, camera monitors are typically used for one of two tasks: broad monitoring of several live camera feeds or scanning archived video files for a specific event, person, or object, usually in the context of an investigation.
Human monitors, on the other hand, soon become cognitively inundated, losing essential content even while watchful (Faber et al.; 2012; Keval and Sasse; 2010). As a result, modern camera systems are hit-or-miss tools for observing and detecting ongoing situations, as well as expensive, time-consuming search platforms for locating specific video sequences (Gill; 1994; Goold; 2004; Hier et al.; 2007; Näsholm et al.; 2014; Ratcliffe et al.; 2009; Sasse; 2010).

1.4 Computer Vision

Computer vision is becoming more popular as a solution to the shortcomings of human-monitored camera networks (CV). The promise of CV-enhanced cameras is that they will eliminate the two most common human drivers of security camera inefficiency. During real-time monitoring, a computer algorithm will not become bored or distracted, and occurrences of interest will be detected more rapidly (Idrees et al. 2018). There has been continuous research into the efficiency of CV software for automatically analyzing huge camera networks (Adams and Ferryman; 2015; Coetzer et al.; 2011; Gong et al.; 2011; Gowsikhaa et al.; 2014; Hesse; 2002).

In the early 2000s, there were calls for CV to be integrated into surveillance camera systems, as well as discussions of prospective applications (Baldwin and Baird; 2001; Barrett et al.; 2005; Thomas and Cook; 2006). However, CV's full potential has yet to be realized, and current applications are mostly focused on facial recognition and license plate readers (Adams and Ferryman; 2015). The practice of purchasing camera systems that require human monitors is still prevalent (Keval and Sasse; 2010; Piza et al.; 2019). Autonomous anomaly detection by computer vision analytics plays a critical role in current intelligent video surveillance systems, not only increasing monitoring efficiency but also reducing the stress on monitoring the process. Video anomaly detection has been investigated for a long time, however because to the complexity of modelling anomalous

occurrences and the scarcity of anomaly data, this problem is still far from being solved (as evidenced by the low accuracy on the UCF-Crime (Sultani et al.; 2019) dataset). Detecting anomaly events requires a knowledge of complicated visual patterns, and some patterns, such as arson, burglary, and stealing, can only be recognized when the model learns long-term temporal relationships and causal reasoning.

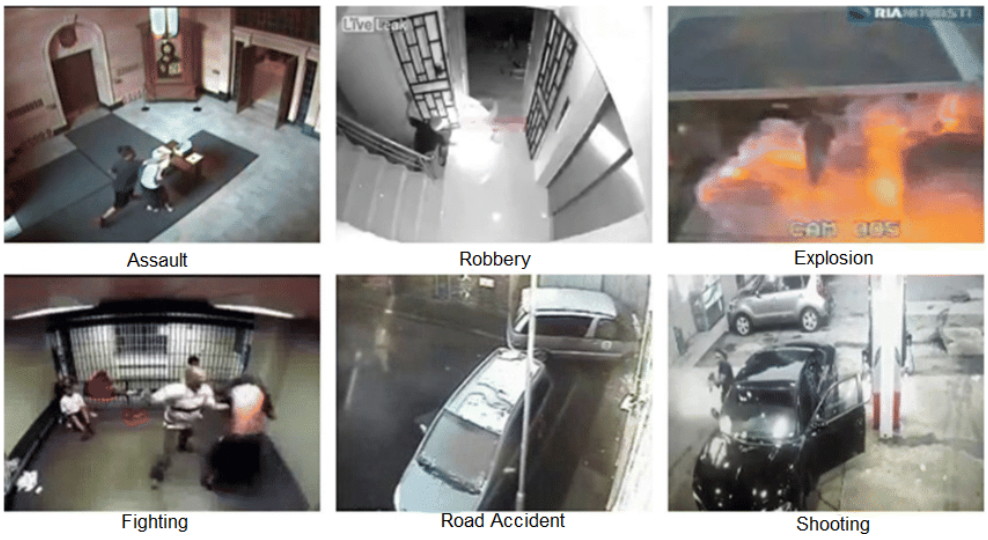


Figure 1: Anomaly Classes

Moving on to the next section of the introduction we introduce the technical approaches used in this research.

1.5 Technical Approach

Recent research has been able to take advantage of large-scale datasets and sophisticated computation resources thanks to deep learning techniques. Following the establishing of unsupervised anomaly detection, a variety of studies based on deep AE (Hasan et al.; 2016; Luo et al.; 2017a,b; Gong et al.; 2019) have been offered (autoencoder).

1.6 Generative Adversarial Networks

Generative Adversarial Networks (GAN) belong to the group of generative models that are way more advantageous over other generative models where these models include the generation of frames in parallel as compared to serial generation in other models that are also generating future frames. This includes the generation of samples in parallel as compared to serial generation in Fully Visible Belief Networks. Generative Adversarial Networks does not require Markov chains when compared to the

Boltzmann machines. It is also observed that Generative Adversarial Networks produces better samples than other models. Generative Adversarial Networks are based on a minimax game where the first model called the Generator directly produces future frames, where the second model called the Discriminator attempts to distinguish between samples drawn from the training data and samples drawn from the Generator model it also takes a fixed length randomly drawn from a Gaussian distribution and is used to seed noise for the generative process.

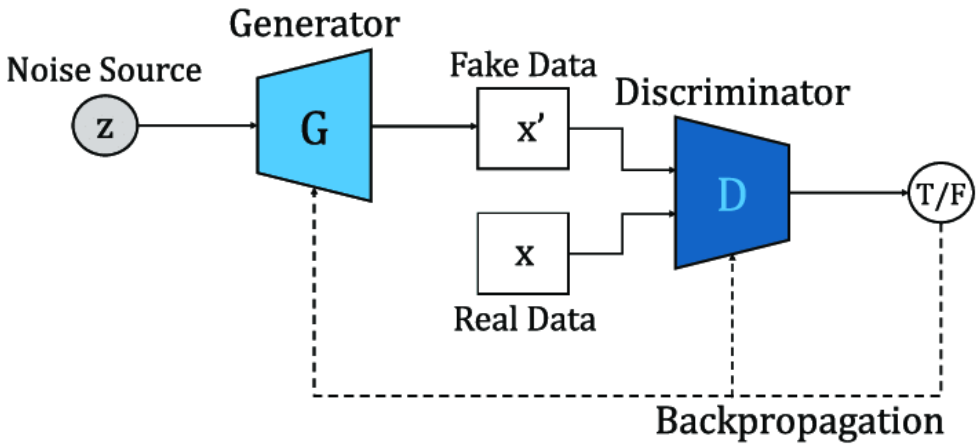


Figure 2: Generative Adversarial Network

1.7 Challenges and Solutions

The challenge of producing future frames from a collection of consecutive frames is known as video prediction, and it has applications in abnormal event detection (Liu et al.; 2018), video coding (Lotter et al.; 2016),video completion, robotics (Finn et al.; 2016), and autonomous driving. This issue has been researched for a long time, and deep learning has lately significantly improved the performance of video prediction algorithms based on deep architecture models such as convolutional neural networks (CNNs) and generative adversarial networks (GANs). Traditional video prediction methods (Patraucean et al.; 2015)compute pixel-wise motion and then estimate the motion of pixels in future frames based on the assumption that motions are linear. This principle is inherited by a number of deep learning-based approaches (Liang et al.; 2017; Tulyakov et al.; 2018; Villegas et al.; 2017) They use deep networks, such as FlowNet (Long et al.; 2015), to explicitly compute pixel-wise motion, and then the motion information is combined with training images to generate future frames. Despite the fact that the concept is identical to the traditional approach, deep networks show promise when dealing with complex motions in a dynamic context. Computing pixel-wise motion is prone to errors due to

lighting changes, occlusion, and rapid camera motion, which is one of the key drawbacks of this approach. Deep networks can predict realistic future images without specifically computing pixel-wise motion, according to a number of research (Jia et al.; 2016; Kalchbrenner et al.; 2017; Mathieu et al.; 2015; Vondrick et al.; 2016; Lotter et al.; 2016) . The bulk of them employ CNNs to predict future frames (Byeon et al.; 2018; Kalchbrenner et al.; 2017; Jia et al.; 2016; Lotter et al.; 2016), however CNN-based approaches frequently produce fuzzy predictions because they try to minimize loss across all training images (Ledig et al.; 2017; Byeon et al.; 2018)used the convolutional long term short memory (ConvLSTM) to record both past and spatial contexts to prevent the hazy artefact, which currently provides the highest performance for a few datasets. GANs, on the other hand, have gotten a lot of interest for predicting future frames (Mathieu et al.; 2015; Vondrick et al.; 2016; Liang et al.; 2017), because they train a discriminator network and a generator net at the same time. Recent research has been able to make use of massive datasets and powerful computing resources. Following the establishing of unsupervised anomaly detection, a variety of studies based on deep AE (Hasan et al.; 2016; Luo et al.; 2017a,b; Gong et al.; 2019) have been offered (autoencoder). Hasan et al. (Hasan et al.; 2016) propose using an FCN (Fully Convolutional Network) based AE to learn both motion features and discriminative regular patterns. The regularity score is calculated using the AE model's reconstruction error.(Luo et al.; 2017a) combines FCN with LSTM (long short-term memory) as a ConvLSTM-AE to better model the temporal relationship inside a video, which enhance the efficiency of the AE framework.(Luo et al.; 2017b) investigates the use of RNNs in conjunction with sparse coding (Recurrent Neural Network). To bring video temporal information into the background of sparse coding, a temporally-coherent sparse coding system is suggested. (Gong et al.; 2019) presents a memory-enhanced AE for anomaly detection that memorizes prototypical normal patterns. The memory is then accessed, and subsequent frames are reconstructed using attention-based sparse addressing. The anomaly events are determined using the reconstruction error in all of the AE-based approaches discussed. On the other hand, (Ionescu et al.; 2019) recommends using k-means clustering and one-versus-all SVM to structure the problem as a multi-class classification (Support Vector Machine). An alternative to directly estimating the reconstruction error of future frames with a set of basis or AE is to anticipate future frames based on past frames and provide a high anomaly score when the real future frame differs significantly from the predicted one. The idea of GANs (Generative Adversarial Networks) is developed to achieve future prediction, where a generator

and a discriminator are trained alternately to achieve opposite aims. The generator tries to create frames that seem like actual frames, while the discriminator is trained to tell the difference between fake and real frames. With enough training data and the right procedures, the generator could make incredibly realistic fake frames that the discriminator couldn't tell apart from the actual ones. To forecast future frames, most recent publications (Liu et al.; 2018; Ye et al.; 2019) use an FCN-based architecture as the generator. Liu et al. (Liu et al.; 2018) propose to include constraint on intensity, gradient and motion for future frame creation. The intensity constraint gives the consistency between generated frames and real ones on RGB space, while the gradient constraint can sharpen the generated images. By minimizing the optical flow difference between predicted and real frames, the motion constraint seeks to generate predicted frames with similar motions to the real ones. Based on the GAN-based approach, Ye et al. (Ye et al.; 2019) also offer a predictive coding module and an error refinement module. Now the generative models have reached a point where it can generate a complete scene through the Generative Adversarial Networks and their advances in the current technical field as researchers have reached a good generative model called the Cycle Generative Adversarial Network making a breakthrough result in generating fake videos and fake frames to produce a lookalike to the real video where the human eye might not distinguish between the fake and real video.

1.8 Our proposed solution

This research proposes a framework integrating two models: a future frame prediction method and a weakly supervised deep learning classification method. The framework was used to predict anomalous behavior in surveillance cameras to prevent crime. They used the future frame prediction (GAN) method to close the gap left by insufficient anomaly prediction methods. By predicting future frames containing anomalous behavior before it happens by a certain duration, they showed the possibility of depending on the predicted frames, as accuracy either increased or did not change using the given datasets. Our results not only open possibilities in anomaly prevention but also provide a good understanding of what could be done with new improvements in future frame generation methods. In future works, we are looking forward to modifying Generative Adversarial Networks to produce a higher number of frames with higher Structured Similarity Index Measure to save as much time as possible to predict future anomalies much earlier and improve the deep learning method to

achieve higher accuracy using the given datasets.

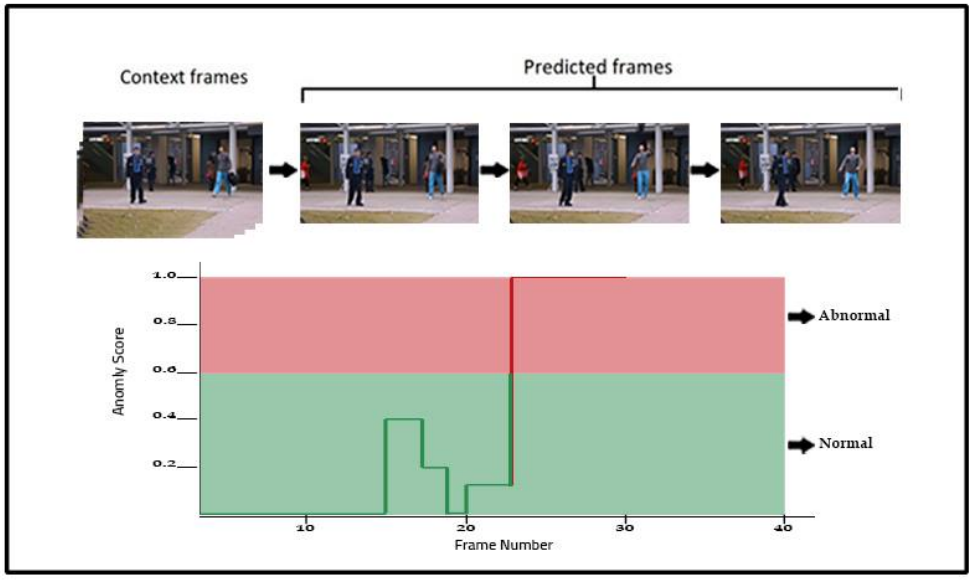


Figure 3: Proposed Model

2 Related Work

In computer vision, detecting aberrant actions is a difficult challenge. For starters, there is no universal definition for anomalous phenomena. The concept of aberrant behavior changes depending on the situation and context. While riding a bike on the street is considered a typical activity, it may be classified as aberrant if it occurs in pedestrian paths. One typical strategy to dealing with this problem is to regard the unseen events as aberrant. However, gathering all typical events is impossible, and there will always be unseen scenes containing normal activity. As a result, the second issue is a lack of labelled data. Detecting and classifying such events is a time-consuming process. Furthermore, dealing with video input is tough since the issue space is a high dimensionless space with a huge number of redundant information, making it difficult to obtain a meaningful feature that can capture relevant information. Another issue is that some abnormal actions are undetectable at a low timescale, which is typical of frame level computation. In this section, they discuss several methods for detecting anomalous events in videos.

2.1 Sparse Reconstruction Cost for Abnormal Event Detection Summary

This paper(Cong et al.; 2011) proposed a method to detect abnormal events using a sparse reconstruction depending on normal videos. Their idea is to get the sparse reconstruction cost (SRC) of the testing videos to measure the normalness and comparing it with normal videos from the dataset, since the generated reconstruction cost coefficients of abnormal events are extremely higher than the generated from normal ones. The paper introduces a selection model to generate inputs of minimal size, discard redundant, and noisy training samples which lead to an increase in computational efficiency. They provide a unified solution to detect both local abnormal events (LAE)and global abnormal events (GAE). Experiments on three benchmark datasets (UMN dataset is used to test the GAE; UCSD dataset and the Subway dataset are used to detect LAE).

2.2 Learning Temporal Regularity in Video Sequences

This paper (Hasan et al.; 2016) presents a model to learn normal motion patterns using auto-encoders. This model consists of two methods. Firstly, take the advantage of the conventional spatial-temporal local features and

learn a fully connected autoencoder, however this stage is may suboptimal for learning temporal regularity as they are not designed or optimized for this problem, so they go directly to the second method by learning both the motion features and classifies the input to normal or abnormal in a single learning framework using a fully convolutional neural network based autoencoder. This model trained on multiple video sources (different datasets: CUHK Avenue, Subway, and UCSD Pedestrian datasets) to make a single model for different videos. The proposed model is computationally more efficient than sparse coding, but generate a few more false alarms because it identifies any deviations from regularity, and many of which have not been annotated as abnormal events in those datasets for example in CUHK Avenue dataset "running" event is classified to abnormal behavior due to it is unusual motion pattern, but in the ground truth it is a normal event. The theory is that the learnt autoencoder will accurately reconstruct motion signatures in regular videos with low error, but will not recreate motions in irregular videos.

2.3 Observe Locally, Infer Globally: a Space-Time MRF for Detecting Abnormal Activities with Incremental Updates

This paper(Kim and Grauman; 2009) is introduced a space-time Markov Random Field (MRF) model to detect abnormal behaviors in video. The proposed model divides a video into a grid of local regions. Each region corresponds to a node, and neighboring nodes are connected with links. Each node associated with optical flow to learn normal patterns of at each node, then Mixture of Probabilistic Principal Component Analyzers (MPPCA) to learn motion patterns to obtain probabilistic estimates of whether each node is normal or abnormal. Besides, the proposed model invented an incremental update for the MPPCA and their parameters whenever new observations come in. The model detects abnormal behaviors in both local and global situations.

2.4 Dominant motion analysis in regular and irregular crowd scenes

Based on corner features, this research(Kim and Grauman; 2009) developed a novel method for dominating motion analysis in cluttered settings. The method consists of three processing stages namely: corner features extraction, corner features snipping with an enthalpy model, and

random forest inferencing. Firstly, extract the corner features from video frames and track them using pyramidal Lucas-Kanade optical flow to get the motion patterns. These corner features are passed to Enthalpy Model, which isolate and filter out the features that do not contribute to the identification of the dominant crowd motion returning only potential interest features only. Finally, the obtained potential interest features are trained on Random Forest to learn the behaviors of crowd leading to detect the dominant crowded motion. This method trained on two benchmark video sequence datasets (UCD, UCF).

2.5 A revisit of sparse coding based anomaly detection in stacked rnn framework

Luo et al. proposed a sparse coding based method that learns a dictionary of normal events (Luo et al.; 2017b). In order to apply temporal coherency between neighboring frames, they propose a temporally coherent sparse coding that preserves the similarity between adjacent frames. In order to achieve this, they define a new objective function. Then they show that their model can be interpreted as a special stacked Recurrent Neural Network (sRNN) and they show how their TSC model parameters can be mapped into sRNN.

2.6 Stochastic Adversarial Video Prediction

(Lee et al.; 2018) introduced a video prediction to Predicting what will happen in the future necessitates a thorough comprehension of the physical and causal laws that govern the universe. From robotic planning to representation learning, a model that can do so has a lot of potential applications. Learning to predict raw future observations, on the other hand, model that combines latent variables trained using a variational lower limit with an adversarial loss to achieve great visual and physical realism. Their method can create a variety of stochastic predictions thanks to VAE-style training using latent variables, and our studies demonstrate that the adversarial loss is effective at providing predictions that are more visually realistic in the eyes of human raters. They test their technique, as well as ablated variants that only have the VAE or simply the GAN loss, on a range of quantitative and qualitative criteria, such as human ratings, diversity, and accuracy of the projected samples. Their findings show that their method generates more realistic predictions than previous methods while maintaining the sample diversity of VAE-based methods.

2.7 Future Frame Prediction of a Video Sequence

In this research (Kaur and Das; 2020)Predicting future frames of a video sequence has been a hot topic in the field of Computer Vision since it offers a wide range of applications. The ability to forecast, anticipate, and reason about future occurrences is at the heart of intelligence, and one of the primary aims of decision-making systems like human-machine interaction, robot navigation, and autonomous driving. The difficulty resides in the ambiguous nature of the problem, as numerous future sequences for the same input video shot may be possible. In a basic model, numerous probable futures are averaged into a single hazy projection. They suggested a unique architecture for the job of video prediction that combines the strengths of the two approaches employed in "Stochastic Adversarial Video Prediction" (Lee et al.; 2018) and "Eidetic 3D LSTM: A Model for Video Prediction and Beyond" (Wang et al.; 2018). They statistically and qualitatively assessed four models: SAVP, E3D-LSTM, Modified SAVP, and our Proposed Architecture on three datasets: UCF101, Moving MNIST, and Penn Action Dataset. They found that our suggested Architecture outperformed the baseline models by a large margin. Models will be tested on synthetically created datasets and alternative assessment metrics will be used in the future.

2.8 Early Action Prediction with Generative Adversarial Networks

In this paper (Wang et al.; 2019) They present the generative adversarial network to address the difficult challenge of early action prediction. The suggested method improves the characteristics of partial videos by communicating discriminative information from corresponding complete movies, which is accomplished via a generator and discriminator network that are alternately tuned. The generator learns to improve partial video features by calculating additive residual error between them and original video features. The discriminator seeks to separate the generator's enhanced features from the original features in entire videos. Furthermore, by back-propagating the action class information to the generator, a perceptual network is used to increase the discriminability of augmented features. The generator provides more discriminative and informative features for partial films as a result of the competition between these networks, boosting action prediction performance. Extensive trials have shown that the suggested strategy is superior in action prediction, particularly early action prediction.

2.9 Future Frame Prediction for Anomaly Detection – A New Baseline

This is the first paper(Liu et al.; 2018) to introduce a method to detect anomaly detection problem within a video prediction framework that calculate the difference between a predicted future frame and its ground truth to detect abnormal behaviors. The idea of the paper is train on normal videos to predict a high-quality future frame for normal behaviors, so in the testing phase if the predicted future frame agrees with its ground truth so it corresponds to a normal behavior otherwise it corresponds to an abnormal behavior. They also introduce for the first-time temporal constraints in the future frame prediction task by enforcing the optical flow between frames beside the (spatial) constraints. This paper used generative adversarial network (GAN) based on U-net network architecture and trained on three datasets (CUHK Avenue, UCSD, ShanghaiTech).

2.10 High Resolution Video Generation using Spatio-Temporal GAN

In this research(Sagar; n.d.) They present a unique network for producing high-resolution video. By applying a k-Lipschitz constraint on the loss term and employing class labels for training and testing, our network borrows ideas from Wasserstein GANs and Conditional GANs. They show the layer wise details of the Generator and Discriminator networks, as well as the combined network architecture, optimization details, and algorithm used in this study. A combination of two loss terms is used in their network: mean square pixel loss and adversarial loss. The UCF101, Golf, and Aeroplane Datasets were utilized to train and test our network. Our network surpasses earlier networks on unsupervised video production using Inception Score and Fréchet Inception Distance as assessment metrics.

2.11 Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning

(Lotter et al.; 2016) proposed a method of deep convolutional recurrent neural network inspired by the principles of predictive coding. The model was trained to predict the next-frame video prediction with the belief that prediction was an effective objective for unsupervised learning.

2.12 Classifying Pedestrian Actions In Advance Using Predicted Video of Urban Driving Scenes

In this research(Gujjar and Vaughan; 2019) They suggested and tested three different types of neural network algorithms for making future video predictions. They next created a Temporal Variation Graph for all models to assess their contributions in terms of per-frame visual reproducibility and temporal coherence. Their findings imply that residual connections enable learned intermediate representations to diverge from one another. Iterative refining is visible with multi-stage recurrent decoding. Their method is unique in that it learns a sequence of representations from an encoder rather than a comprehensive vector, as many other sequence generation methods do. They also devised and tested a C3D action classifier model-based classifier algorithm. The network was given the goal of recognizing a crossing action by looking at a video of a future event and so predicting a pedestrian's crossing intent.

2.13 Anticipating Pedestrians Crossing With Future Frames Prediction

In this research(Chaabane et al.; 2020a) They provide an end-to-end future-prediction model focusing on pedestrian safety in this work. Their programme, in particular, leverages earlier vi deo frames taken from the vehicle's perspective to forecast whether a pedestrian will walk in front of it. The long-term goal of this research is to create a fully autonomous system that acts and reacts in the same way that a defensive human driver would, anticipating future occurrences and reacting to reduce danger. Because of the substantial danger of harm to pedestrians if their activities are mis predicted, they concentrate on pedestrian-vehicle interactions. Their end-to-end model is divided into two stages: the first is an encoder/decoder network that learns to predict future video frames, and the second is an encoder/decoder network that learns to predict future video frames. The second stage is a deep spatiotemporal network that uses the first stage's anticipated frames to forecast the pedestrian's future behavior. On the Joint Attention for Autonomous Driving (JAAD) dataset, their system obtains high accuracy in predicting pedestrian behavior and future frames.

2.14 Deep Reinforcement Learning for Real-world Anomaly Detection in Surveillance Videos.

In this research(Aberkane and Elarbi; 2019) Surveillance videos are regarded as a critical component of every smart city initiative. Deep learning has been integrated with reinforcement learning techniques in recent years to build usable representations for situations with high dimensional raw data input. They construct a Deep Q Learning Network (DQN) in this research to pinpoint anomalies in videos by teaching the agent how to detect and recognize irregularities in films. Multiple instance learning (MIL) strategies based on common share features with reinforcement learning inspired their concept. They think of typical and anomalous videos as bags, and video clip selection as action. They created a fully connected layer in their DQN architecture that computes probability for each video segment in both positive (anomalous) and negative (normal) bags, indicating how probable a clip is to include an anomaly. Their method is used on the UCF-Anomaly-Detection-Dataset, a new large-scale dataset comprising 128 hours of films that consists of 1900 lengthy and untrimmed real-world surveillance recordings with 13 occurrences of realistic anomalies.

2.15 Long-Term Human Video Generation of Multiple Futures Using Poses.

In this research(Fushishita et al.; 2020) Predicting human behavior in the near future from an observed human video is a crucial task for a range of applications (e.g., robotics, autonomous autos). They describe a novel way for generating long-term future movies of numerous futures from an input human video utilizing a hierarchical strategy that involves first predicting future human poses and then generating the future video in this paper. They present a new network that uses a unidimensional convolutional neural network in adversarial training to predict long-term future human pose sequences. They also offer two more inputs: a hidden code and an attraction point, which may be used to forecast a range of multiple futures. Finally, the videos created using our predicted positions are long and varied. The suggested method outperforms the others in terms of realism, diversity, and correctness of generated postures and videos, according to experimental data. Because their technology makes videos frame by frame, they may use the most recent image to generate videos with a greater resolution in the future.

2.16 Future video generation using human pose

The use of a human pose input is one of the most successful ways for producing human video. Yan et al. used an input frame and a sequence of future human poses to create future video. Villegas et al. used an LSTM to predict future human positions as body-joint coordinates, then created video frame by frame based on the projected poses. This method succeeded in producing long-term films, but it is unable to produce numerous futures because an LSTM's output does not vary for the same input. Cai et al. introduced a network that generates human pose sequences using latent noise and an action class label, as well as another network that generates video from the created poses. This model can be expanded to produce a future pose sequence from a past pose sequence, but not many futures. Furthermore, because the action class of the input movement is not accessible, applying an action class label is inadequate for future prediction. Walker et al. used a combination of an LSTM and a vibrational auto encoder (VAE) to produce numerous human positions from two pose sequence inputs, and then used 3D convolutional neural networks to generate a video. The VAE allows several human poses to be generated, which are then input into the LSTM to anticipate a sequence of future poses. However, because LSTM mistakes compound exponentially, this approach is inappropriate for long-term future prediction. They present a strategy for long-term video prediction of several futures in this research. They use unidimensional convolutional neural networks to produce long-term near-future sequences, which allow them to generate sequences without vanishing gradients or error propagation issues. They then use two criteria to encourage our network to build a range of possible futures: a latent coding that stimulates a specific form of motion, and an attraction point that induces motion towards a specific spot in the image.

2.17 Deep Convolutional Generative Adversarial Networks Based Flame Detection in Video

This paper(Aslan et al.; 2019) proposed a real time model for fire detection using Deep convolutional network GAN (DCGAN) in surveillance cameras. The model training is composed of two stages. Firstly, they train the DCGAN using data containing fire. So, the discriminator of the DCGAN learns the representation of fire in the images and distinguish the non-fire images. Secondly, they train the discriminator without the generator on non-fire videos obtained from surveillance cameras. The second stage makes the model more robust. The results show that the model detects

fire effectively with low false alarms rates in real time.

2.18 Video Anomaly Detection Via Predictive Autoencoder with Gradient-Based Attention

In this paper(Lai et al.; 2020) they present a novel two-branch predictive autoencoder, consisting of a reconstruction decoder and a prediction decoder, in which the prediction decoder generates future frames and performs anomaly detection by making a distinction between projected future frames and their ground truth. And the reconstruction decoder reconstructs the current frame, which can constrain the encoder to learn video representations better. Furthermore, the reconstruction decoder offers gradient-based concentration, which aids the prediction decoder in producing higher-quality future frames. Their method unifies reconstruction and prediction methods in an end-to-end framework, and on certain publicly available datasets, it achieves amazing results with better projected future frames.

2.19 Anomaly Detection of Predicted Frames Based on U-Net Feature Vector Reconstruction

In this research(Qiang et al.; 2020) they utilize an unsupervised training strategy to improve the anomaly detection system, the reconstruction of possible characteristics of the anticipated frame and u-net actual truth. They lower the reconstruction error between u-potential net's features in the projected frame and the real frame's viable features. The reconstruction error of the full anticipated frame is then reduced using various restrictions, according to the generative adversarial training. When aberrant conduct is discovered, the reconstruction error value surpasses the defined threshold to determine whether abnormal behavior occurred in the surveillance video, thanks to the usage of normal behavior sample training.

2.20 Abnormal Event Detection in Videos using Generative Adversarial Nets

This paper(Ravanbakhsh et al.; 2017) introduced a method for anomaly detection in the crowd using a generative adversarial network based on future frame generation and optical flow. They used GANs to learn the normality of human behavior. They trained two networks, the first one is for generating optical flow and the second one is for generating a frame

from optical flow. Their network is based on conditional generator and discriminator. They generated the models on only normal videos. At testing time, they used the generator to produce appearance and motion. As the generator trained in normal behavior only, the generator generates appearance and motion of normal action and then calculates the difference between the generated frame and the real frame to detect anomalies. They computed their accuracy based on two-way first-and-second-frame level and pixel level on Ped1 and Ped2.

3 Datasets

3.1 1M Sports

The Sports-1M is a collection of over a million YouTube videos. The authors provided a YouTube URL that can be used to access the videos in the dataset. Unfortunately, since the dataset was generated, around 7% of the films have been removed by YouTube uploaders. This may affect the training, validation, and/or testing sets utilized in various research. However, the collection still contains over a million recordings, divided into 487 sports-related categories with 1,000 to 3,000 videos each. By examining the text metadata connected with the videos, the YouTube Topics API is used to automatically categories the videos with 487 sports classes (e.g. tags, descriptions). While a large-scale dataset like Sports-1M may be useful for training CNN-based algorithms that are susceptible to overfitting on smaller datasets like UCF101 and HMDB51, it should be used with caution. For starters, because videos are collected automatically, labels are limited. Second, around 5% of the films are labelled with multiple classes. As a result, the training film may fail to depict distinguishing characteristics of individual acts. Third, because YouTube users might upload duplicate videos, the same video may appear in both the training and testing sets. The videos have a spatial resolution of 400x240 to 1280x720 pixels and a runtime of 0 to 37,427 frames. The Sports-1M dataset is divided into three sections: 70% training, 10% validation, and 20% testing. The videos should be tested using a 10-fold cross-validation method. We divided the dataset into three parts: a training set with 70% of the videos, a validation set with 10% of the videos, and a test set with 20% of the videos. Due to the possibility of identical videos on YouTube, the same video may appear in both the training and test sets. To gain a sense of the scope of the problem, we analyzed all videos on a frame-by-frame basis with a near-duplicate detecting algorithm and discovered that only 1755 videos (out of 1 million) include a significant fraction of near-duplicate frames. Moreover, because we only use a random collection of up to 100 half-second clips from each video and our videos are on average 5 minutes and 36 seconds long, the same frames are unlikely to appear across data splits.

3.2 Avenue Dataset

CUHK Avenue dataset contains 16 training videos and 21 testing ones with a total of 47 abnormal events, including throwing objects, loitering

and running. The size of people may change because of the camera position and angle.

3.3 ShanghaiTech Dataset

ShanghaiTech dataset is collected in ShanghaiTech University under 13 scenes with complex light conditions and camera viewpoints. It consists of 437 videos with 726 average frames each. The training set consists of 330 normal videos and testing set contains 107 videos with 130 anomalies. Anomaly events include unusual patterns in campus such as bikers or cars.

3.4 UCF Crime

UCF Crime (Sultani et al.; 2019) is a collection of 1900 uncut movies that depict 13 real-life abnormal incidents, including Abuse, Arson, Assault, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. There are 950 regular videos and at least one abnormal occurrence in each of the remaining recordings. There are 800 regular videos and 810 anomalous videos in the training set. For validation, the remaining 150 normal and 140 abnormal videos are chronologically tagged. All 13 anomalous events are covered in both the training and testing sets. Some movies may have numerous oddity categories, such as robbery and fighting, burglary and vandalism, and arrest and gunfire. All of the videos are realistic enough for real-world monitoring. Furthermore, UCF Crime is extremely difficult since it includes a wide range of light conditions, image resolutions, and camera postures in complex settings.

3.5 UCSD

A stationary camera situated at an elevation viewing pedestrian pathways was used to collect the UCSD Anomaly Detection Dataset. The number of people in the pathways varied, from low to quite packed. The video contains simply pedestrians in its natural state. Bikers, skaters, small carts, and people walking across a walkway or in the grass that surrounds it are examples of common anomalies. There were a few instances of people in wheelchairs as well. All abnormalities were not staged for the purposes of assembling the dataset; they occurred naturally. The data was divided into two groups, each of which corresponded to a different scene. Each scene's video footage was divided into several clips of around 200 frames each. The UCSD dataset is divided into two parts, Ped1 and Ped2. They were photographed in two locations on the UCSD campus

where the majority of pedestrians walk. Only normal frames are used in the training set (34 clips for Ped1 and 16 clips for Ped2), while both normal and anomaly frames are used in the test set (36 clips for Ped1 and 12 clips for Ped2). All test clips have frame-level annotation, and 10 of them have pixel-level ground-truth. Non-pedestrian entities such as bikers and skaters are defined as anomaly instances in the UCSD dataset because pedestrians walking is considered the normal pattern. The UCSD dataset contains two parts: The UCSD Pedestrian 1 (Ped1) dataset and the UCSD Pedestrian 2 (Ped2) dataset. The UCSD Pedestrian 1 (Ped1) dataset includes 34 training videos and 36 testing ones with 40 irregular events. All of these abnormal cases are about vehicles such as bicycles and cars. The UCSD Pedestrian 2 (Ped2) dataset contains 16 training videos and 12 testing videos with 12 abnormal events. The definition of anomaly for Ped2 is the same with Ped1. Usually different methods are evaluated on these two parts separately.

3.6 Subway

The Subway dataset(Adam et al.; 2008) is divided into two parts: Subway Entrance and Subway Exit. In each subway station, there is only one long surveillance video. They were first proposed for real-time detection of unusual events in crowded subway scenes, such as moving in the wrong direction or not being paid.

3.7 UMN

The UMN(of Minnesota; 2019) dataset (University of Minnesota) consists of five videos shot from various perspectives. Walking is the normal pattern, and running is the main anomaly activity.

3.8 DAD

The DAD(Chan et al.; 2016) (Dashcam Accident Dataset) is a proposed algorithm for detecting accidents. Vehicles move around in the normal pattern, and anomaly events include traffic accidents such as car-to-car collisions or motorbike-to-motorbike collisions. The DAD dataset contains 678 videos from six different cities. For training, 58 videos are used. For the remaining 620 videos, 1130 normal clips are sampled as positive clips and 620 clips with accidents are sampled as negative clips. The clips are then divided into two subsets at random: 455 positive and 829 negative clips for training, and 165 positive and 301 negative clips for testing.

3.9 CADP

CADP(Shah et al.; 2018) (Car Accident Detection and Prediction) is a research project that focuses on car accidents captured by CCTV (Closed-Circuit Television) cameras. CADP's 1416 videos contain traffic accidents, with 205 of them having both temporal and spatial annotations. CADP contains videos shot with a variety of camera types, qualities, and weather conditions, and the anomaly events are realistic enough for real-world use.

3.10 A3D

A3D(Yao et al.; 2019) is a collection of 1500 dashboard camera video clips of on-road abnormal events. Human annotators have annotated the start and end times of each abnormal traffic event in each video. At 10 frames per second, a total of 128,175 frames (ranging from 23 to 208 frames) are clustered into 18 types of traffic accidents.

3.11 DADA

DADA(Fang et al.; 2019) is a traffic accident dataset that was gathered to predict driver attention in accidental scenarios. It contains 658,476 available frames in 2000 videos with a resolution of 1584x660 pixels. Based on the participants of accidents, the videos are divided into 54 different categories, such as "hitting" and "out of control" (e.g. pedestrian, vehicle, cyclist, etc.). The spatial crash-objects, as well as the temporal window of accident occurrence, are annotated.

3.12 DoTA

The DoTA (Yao et al.; 2020) (Detection of Traffic Anomaly) dataset contains 4,677 videos with temporal, spatial, and categorical annotations. The goal is to develop a when-where-what pipeline for detecting, locating, and recognising anomalous events in egocentric videos. The video clips were gathered from YouTube channels that featured a variety of dash camera accident videos from various locations. The video clips were culled from YouTube channels that featured a variety of dash camera accident videos from various countries, shot in a variety of weather and lighting conditions.

3.13 DOT Traffic

The Iowa DOT (Department of Transportation) Traffic dataset (Naphade et al.; 2019) consists of 200 videos, each about 15 minutes long and recorded at 30 frames per second with a resolution of 800 410 pixels. Each training and testing set includes 100 videos. It does not provide annotation for the testing set as the official dataset for the 2018 AI City challenge(Naphade et al.; 2019) Track 3. Car accidents and stalled vehicles are the most common anomaly patterns.

Table 1: Datasets

DATASET	# OF VIDEOS	AVERAGE FRAMES	QUALITY	FUTURE FRAME APPLICATION	COMPUTATION PROBLEM
UCSD PED1	70	201	LOW	✓	✓
UCSD PED2	28	163	LOW	✓	✓
SUBWAY ENTRANCE	1	121,749	LOW	✗	✓
SUBWAY EXIT	1	64,901	LOW	✗	✓
AVENUE	37	839	HIGH	✓	✓
UMN	5	1,290	MEDIUM	✗	✓
DAD	1,730	100	MEDIUM	✗	✗
CADP	1,416	366	MEDIUM	✗	✗
A3D	1500	85	LOW	✓	✗
DADA	2,000	324	LOW	✗	✗
DOTA	4,677	156	MEDIUM	✓	✗
IOWA DOT	200	27,000	LOW	✗	✗
SHANGHAITECH	437	726	HIGH	✓	✓
UCF CRIME	1,900	7,247	LOW	✓	✗
SPORTS 1M	1,000,000+	N/A	MEDIUM	✓	✗

4 Methodology

A few recent studies have mentioned the possibility of predicting actions that are future frame–dependent in other fields, such as autonomous driving and action recognition Chaabane et al. (2020b); Gujjar and Vaughan (2019), which drives our work to predict anomalous behaviours before they happen. The proposed method focuses on integrating two main methods, which include the future frame prediction method called retrospective cycle GAN Kwon and Park (2019) and the weakly supervised detection method MIL Sultani et al. (2019). They are both implemented in a single framework installed on surveillance cameras to predict future anomalies depending on the future frame extracted from the GAN model. Each frame undergoes some preprocessing to be compatible with the retrospective cycle GAN and is then switched into a sequence of frames, which is taken as an input to the retrospective cycle GAN to predict future frames. These frames then undergo feature extraction using a pretrained feature extraction model; the features extracted are then fed into the MIL model, which then produces a ranking score that is directly proportional to the anomaly. Hence, if the anomaly's value in a frame is high, the ranking score is high, according to a threshold that corresponds to the data's nature.

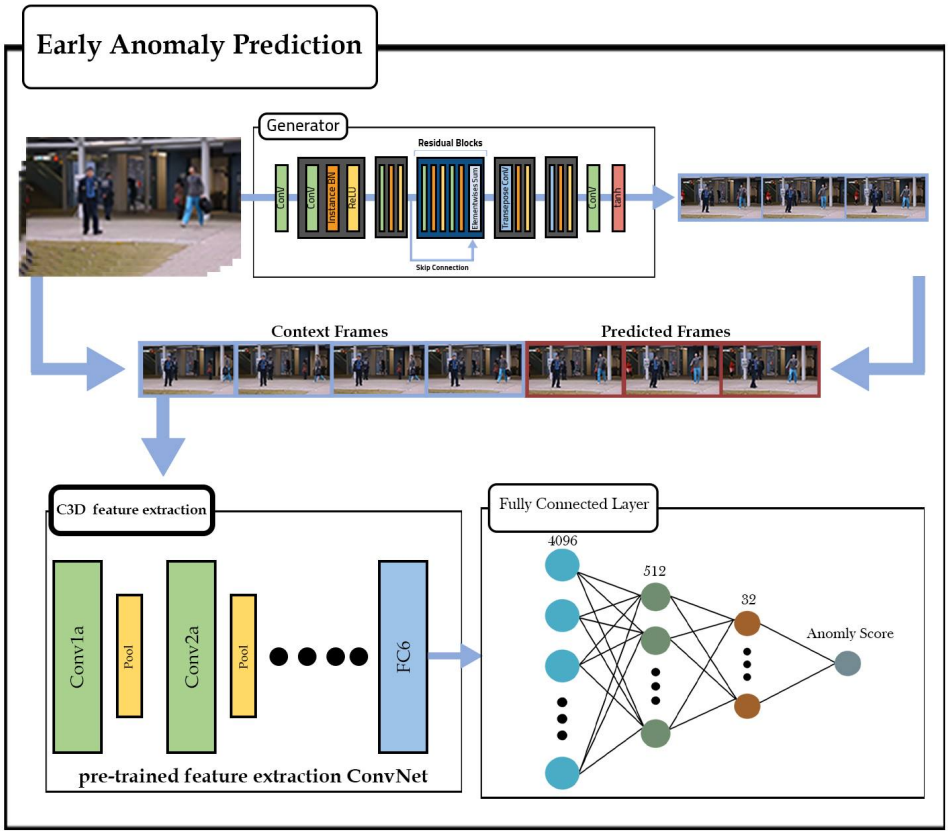


Figure 4: The integration of the models

5 Design Specification

Our design specification is split into two parts the first part consists of the first method which is the anomaly detection and the second part consists of the second method which is the future frame prediction, where these models work together to produce our pipeline, this pipeline is used to predict future anomaly prediction that is used in anomaly prevention.

5.1 Future Frame

Through a very sophisticated research we found that there are many papers that produce promising results, researchers now compete to produce the highest results possible for the future frame prediction methods to produce more frames and higher Structured Similarity Index Measure (SSIM), starting with the Deep Neural Networks such as the PredNet(2016) till the era of generative models such as the Generative Adversarial Networks(2018) which makes the future frame prediction is one of the most challenging problems to have ever existed, researchers look for methods to produce a higher number of frames with high Structured Similarity Index Measure to predict either anomalous behaviors or actions depending on the dataset. Future Frames are produced easily in the action datasets such as UCF-101 and Sports-1M, as it predicts the future frame in a still environment (e.g MoCoGAN), while the anomalous behaviors are not necessarily in still environment as some anomalous behaviors involve explosions and car crashes/accidents which will result in the movement of the camera or the shaking of the camera. The first research to have mentioned the future frame prediction is the Liu paper(Future Frame Prediction for Anomaly Detection – A New Baseline) which shows the ability of reconstructing the frame from scratch and compare it to the ground truth in order to extract the anomalies in the video, but they only reconstruct one frame in order to "detect" the anomaly which is not sufficient, they did not focus on producing a group of frames and their model only focused on a single frame to compare it to the ground truth only. After a grueling research we have found the Retrospective Cycle GAN where it produces a group of frames with an acceptable Structured Similarity Index Measure, this research produces the best results in the current works, they have worked on all the anomalous datasets such as Avenue Dataset.

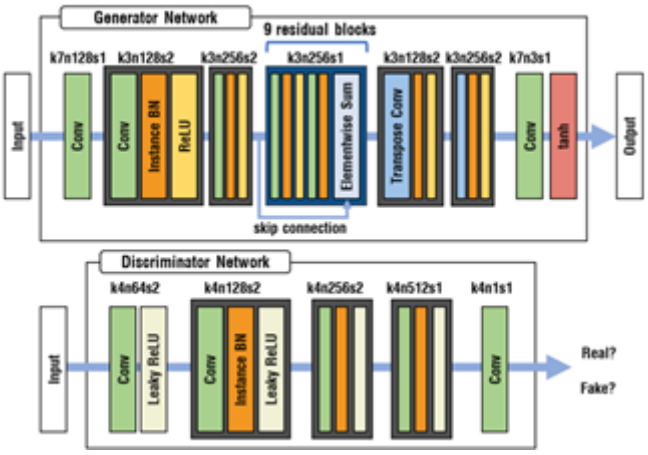


Figure 5: Retrospective Cycle Generative Adversarial Network

5.1.1 Future Frame Prediction

The model is composed of a generator and two discriminators (sequence discriminator and frame discriminator. The generator is used to predict both past and future frames as it takes forward and backward sequences inputs during the training stage enforcing the consistency of bi-directional prediction through the Retrospective cycle constraints. the input sequence may contain fake or real frames. The above stages improve the future frame prediction performance. Moreover, the frame discriminator distinguishes the fake frames from real ones individually while the sequence discriminator judges the output sequence from the generator whether the sequence contains fake frames or not. The sequence discriminator leads to increase in the robustness and temporal consistency of predicted future frame.

5.1.2 Network architecture

The generator and discriminator networks architecture are illustrated in the figure. The generator network architecture consists of 4 convolution layers, 9 residual blocks, and 2 transpose convolution layers. The discriminator network architecture consists of 5 convolution layers with leaky rectified linear units. The two discriminators (frame and sequence) have the same network architecture except the number of input images.

5.1.3 Objective function

In the training of the Retrospective cycle GAN, the objective function consists of four loss functions divided into two categories: two reconstruction losses and two adversarial losses. The objective function also contains three non-zero weights loss functions ($\lambda 1$, $\lambda 2$, and $\lambda 3$) in different parts

used for balancing the train process.

$$L = L_{\text{image}} + \lambda_1 L_{LoG} + \lambda_2 L_{\text{adv}}^{\text{frame}} + \lambda_3 L_{\text{adv}}^{\text{seq}} \quad (5)$$

5.1.4 Notations for explanations

We use these notations in the rest of the paper.

- 1. Denoting the generator as G, frame discriminator as DA and sequence discriminator as DB
- 2. Denoting the input sequence as:

$$\mathcal{X}_{m:n} = \{x_m, x_{m+1}, ..., x_{n-1}, x_n\} \quad \text{s.t.} \quad m < n, \quad (1)$$

  Where m and n are denoting the first and last frame.

- 3.Denoting the target predicted future frame as xn+1 , the predicted future frame from the generator (fake frame) as x/n+1 and the past frame as xm-1.
- 4.Denoting the reversed input sequence as:

$$\bar{\mathcal{X}}_{m:n} = \{x_n, x_{n-1}, ..., x_{m+1}, x_m\} \quad \text{s.t.} \quad m < n. \quad (2)$$

- 5. Denoting the input sequence containing fake frame as:

$$\mathcal{X}_{m:n}^{f} = \{x_{m:n-1} \cup x'_n\},$$

- 6. Denoting the reversed input sequence containing fake frame as:

$$\bar{\mathcal{X}}_{m:n}^{f} = \{\bar{x}_{m+1:n} \cup x'_m\}.$$

- 7.Denoting the predicted future frame when the input sequence contains fake frames as xn+1 to be distinguished from the predicted future frames without fake frames.

5.1.5 Reconstruction losses

The two reconstruction loss functions are used to train the generator. The first reconstruction loss function is formulated by:

$$L_{\text{image}} = \sum_{(p,q)\in\mathcal{S}^{\text{pair}}_{m,n}} l_1(p,q),$$

Where l1(p,q) represents the loss function error between two images. The first loss function minimizes the loss error of image reconstruction for six different pair of images which are:

$$\mathcal{S}^{\text{pair}}_{m,n} = \{(x_m, x'_m), (x_m, x''_m), (x'_m, x''_m), (x_{n+1}, x'_{n+1}), (x_{n+1}, x''_{n+1}), (x'_{n+1}, x''_{n+1})\}.$$

(xn+1, xn+1) and (xm, xm) are used to minimize the prediction errors in both directions (forward and backward) and thus we cannot only predict the future frame, but also predict the previous frame.

$$\mathcal{S}^{\text{pair}}_{m,n} = \{(x_m, x'_m), (x_m, x''_m), (x'_m, x''_m), (x_{n+1}, x'_{n+1}), (x_{n+1}, x''_{n+1}), (x'_{n+1}, x''_{n+1})\}.$$

We also compute the error between (x n+1, xn+1) and (x m, xm), since (xn+1 and xm) are used to predict(xm and xn+1) respectively under one condition that the predicted image(xn+1 and xm) are realistic, so the generator can take this images in the input sequence to predict (xm and xn+1). Moreover, (xm, xm) and (xn+1, xn+1) are used to represent cyclic constraints, since xm and xn+1 are predicted by forward sequence and xm and xn+1 are predicted by backward. Therefore, it can be said that the entire loss function is cyclic and retrospective. Similarly define the second reconstruction function as follows:

$$L_{LoG} = \sum_{(p,q)\in\mathcal{S}^{\text{pair}}_{m,n}} l_1(LoG(p), LoG(q)).$$

The main difference between this loss function and first one is the calculated loss error is after applying Laplacian of Gaussian (LOG) process to detect edges in the input images which removes low frequency information

and high frequency noise. The main purpose of using LOG process is to focus on the structural similarity between images and removing the noise.

The Laplacian of Gaussian is composed two-step process. In the first step, we use Gaussian smoothing filter which is a 2D convolution operator to remove noise from the input image and set a threshold value to distinguish noise from edges. So, if the second derivative magnitude at a pixel after applying the first step exceeds this threshold, the pixel is considered a part of an edge. The second step, we find the zero crossing Laplacian giving better edge localization.

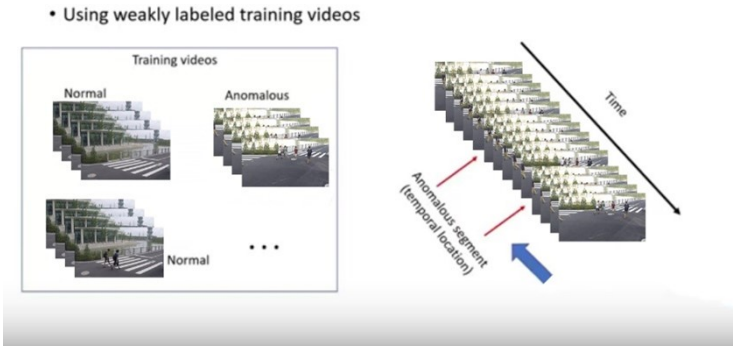
5.2 Anomaly Detection

In our research in the anomaly detection problem, in experiment 1 we noticed that most of the datasets have no annotation, this drove us to search for an approach that is un-supervised or weakly supervised. The weakly supervised issue is formulated as MIL (multiple instance learning). A typical video should have every frame be normal, whereas an anomalous video should include at least one anomalous frame. (Cui et al.; 2011) provides a graph-based MIL framework with anchor dictionary learning, with all experiments taking place on the UCSD (Ding et al.; 2015) dataset in a weakly supervised context. presents UCF-Crime (Joachims; 2002), a deep learning-based technique that includes a large-scale dataset with actual crime-related anomalies and surveillance videos. To extract spatial-temporal information and calculate an anomaly score, a C3D framework is employed. The loss function compels the highest score of a negative video to be higher than the maximum score of a normal video to identify normal and anomaly frames with this poor supervision. On the UCF-Crime dataset, (Joachims; 2002) beats previous works by a substantial margin when the parameters of the C3D model are fixed. The anomaly detection method begins with dividing surveillance videos into a fixed number of segments during training. These segments make instances in a bag by using the positive bags and the negative bags then the features are extracted using the C3D Model pre-trained on the Sports-1M Dataset, we train the anomaly detection model using deep MIL scoring loss function.

5.2.1 Weakly Supervised Learning

Weakly supervised learning is a method that works as the videos has a part of the annotation not a full annotation. However, the annotation in this method is only on video label and not in the temporal realm, which means that the video is labeled as anomaly or not, it will take a high cost

to get people to annotate the data.



We use in this research the incomplete supervision as the videos are weakly labeled as the video is mentioned as anomaly or not only and not specifying the exact temporal annotation of the occurrence of the anomaly we also use the Support Vector Machine (SVM) which is evolved rapidly into the Deep Ranking Loss functions that we have using the Multi-Instance Learning.

5.2.2 Multiple Instance Learning

With the Multiple Instance Learning formulation, a patch-based classifier can be trained using only image-level annotations, with patch level predictions aggregated into image-level scores. Multiple Instance Learning is an algorithm at which works with the weakly supervised technique in which training data is organized into bags, each bag containing a set of instances X = x1, x2,..., xM, and each bag contains only one label Y, Y 0 or 1 in the case of a binary classification problem. Individual labels y1, y2,..., yM are assumed to exist for instances within a bag, but they are unknown during training. A bag is considered negative in the standard Multiple Instance assumption (SMI) if all of its instances are negative. A bag, on the other hand, is positive if it contains at least one positive instance Chandola et al. (2009). The MIL formulation has been widely used to solve high-resolution image classification problems. An image (bag) is divided into M patches (instances), and the classifier treats the patches that belong to the same image together. A positive image (Y = 1) will have at least one positive patch (ym = 1 at least for one m). If the image is negative (Y = 0), then all of its patches are negative (ym = 0 for all m). The max operator can then be used to combine patch predictions to produce an image-wise score: Y = max m(ym), where ym is the patch m prediction. When using this aggregating function, the network's weights will be updated with only one patch per image's information. Other less strict aggregating algorithms have been presented in the literature (Cui

et al.; 2011; Datta et al.; 2002; Arandjelovic et al.; 2016), which use an aggregation of more than one patch prediction per image rather than just the best scoring patch. The labels of all positive and negative samples are available in normal supervised classification tasks using support vector machine (SVM), and the classifier is taught using the following optimization function:

$$\min_{\mathbf{w}} \frac{1}{k}\sum_{i=1}^{k} \overbrace{max(0, 1 - y_i(\mathbf{w}.\phi(x) - b))}^{①} + \frac{1}{2}\|\mathbf{w}\|^2, \quad (1)$$

MIL does not need the temporal annotations which are present in the timeline of the video. In MIL, the temporal locations of events in the videos are unknown. But, just video-level labels identifying the presence of a video abnormality are required. A positive video is one that contains anomalies, while a negative video is one that does not contain any anomalies. Then we express a positive video as a positive bag Ba, in which different temporal segments create individual instances in the bag (p 1, p2,..., pe), where e is the number of instances in the bag. We presume that the anomaly exists in at least one of these circumstances. A negative bag, Bn, is used to represent negative video, with temporal segments in this bag forming negative instances (n 1 , n2 , . . . , nm). There is no oddity in any of the occurrences in the negative bag. Because the precise information (i.e. instance-level label) of the positive instances is unknown, the objective function can be optimized with regard to the highest-scoring instance in each bag Andrews et al. (2002):

$$\min_{\mathbf{w}} \frac{1}{z}\sum_{j=1}^{z} max(0, 1 - Y_{\mathcal{B}_j}(\max_{i \in \mathcal{B}_j}(\mathbf{w}.\phi(x_i)) - b)) + \frac{1}{2}\|\mathbf{w}\|^2, \quad (2)$$

where YBj stands for bag-level label, z stands for total amount of bags, and the rest of the variables are the same as in Eq. 1.

5.2.3 Deep MIL Ranking Model

Anomalous behavior is difficult to define precisely Chandola et al. (2009) because it is highly subjective and varies greatly from one person to the next. Furthermore, how to assign 1/0 labels to anomalies is not obvious. Furthermore, because there are not enough examples of anomaly, anomaly detection is frequently considered as a low-probability pattern recognition problem rather than a classification problem.(Cui et al.; 2011; Sultani

et al.; 2019). Anomaly detection is given as a regression problem in our proposed approach. The abnormal video segments should have greater anomaly scores than the typical video portions. The easiest approach would be to apply a ranking loss that favors high scores for unusual video portions over conventional ones, such as:

$$f(\mathcal{V}_a) > f(\mathcal{V}_n), \tag{3}$$

where Vn and Va are abnormal and normal video clips, and f(Vn) and f(Va) are the available abnormal scores ranging [0,1]. If the segment-level annotations are known during training, the aforementioned ranking mechanism should work well. However, Eq. 3 cannot be used in the absence of video segment level annotations. Rather, we propose the following multiple instance ranking objective function:

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i), \tag{4}$$

where the biggest value is applied to all video clips in each bag. Rather than ranking every case of the bag, we only rank the two examples with the greatest anomaly score in the positive and negative bags, respectively. The real positive occurrence is most likely the section with the highest anomaly score in the positive bag (anomalous segment). The segment with the greatest anomaly score in the negative bag resembles an anomalous segment the most, yet it is actually a typical case. This negative instance is seen as a difficult case that could result in a false alarm in anomaly detection. We aim to push the positive and negative occurrences far apart in terms of anomaly score by utilizing Eq. 4. As a result, our ranking loss in the hinge-loss formulation is as follows:

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i)). \tag{5}$$

One drawback of the above loss is that it ignores the abnormal video's underlying temporal structure. First, anomaly frequently occurs just for a short period of time in real-world circumstances. The scores of the instances (segments) in the anomalous bag should be sparse in this situation, indicating that the anomaly may be contained in only a few segments. Second, because the film is made up of chunks, the anomaly score should fluctuate smoothly between them. As a result, by minimizing the difference

in scores for neighboring video segments, we impose temporal smoothness between anomaly scores of temporally adjacent video segments. The loss function is transformed by including the sparsity and smoothness restrictions on the instance scores.

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$
$$+\lambda_1 \overbrace{\sum_{i}^{(n-1)} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2}^{①} + \lambda_2 \overbrace{\sum_{i}^{n} f(\mathcal{V}_a^i)}^{②}, \quad (6)$$

where 1 presents the sparsity term and 2 presents the temporal smoothness term. The fault is backpropagated from the maximum scored video segments in both positive and negative bags in this MIL ranking loss. We expect the network to learn a generalized model to predict high scores for anomalous regions in positive bags after training on a large number of positive and negative bags Finally, our complete objective function is given by:

$$\mathcal{L}(\mathcal{W}) = l(\mathcal{B}_a, \mathcal{B}_n) + \lambda_3 \left\| \mathcal{W} \right\|_F, \quad (7)$$

where W annotates model weight.

5.2.4 Bags Formations

Each video is divided into an equal number of non-overlapping temporal segments, which we use as bag instances. We extract the 3D convolution features from each video segment (Tran et al.; 2015). We employ this feature representation in video action identification because of its computing efficiency and ability to capture appearance and motion dynamics.

6 Implementation

6.1 Tools

- FFMPEG is a free and open-source multimedia framework that aims to provide the best technical solution for application developers and end users., it able to decode and encode and able to play anything created by humans or machines, it supports the formats from the ancient formats up to the current formats. FFmpeg containing decompressing algorithms and software implementations of videos and audios compressing, that can compiled and run in diverse instruction sets. FFmpeg libraries are core part of software media players, it is including in core processing for YouTube and iTunes. The name of its projects is inspired by MPEG extension. Ffmpeg has three tools:

  1. Ffmpeg it is a command line tool that use to convert multimedia file formats to help in arbitrary sample rates and resize the video, by get the input file and pass it to the demuxer to get encoded data packets then pass it to decoder to get decoded frames then encoder to get encoded data packets then muxer to get the output file.
  2. Ffplay it is a simple media play that based on SDL and MPEG libraries, that use as a testbed for various Ffmpeg APIs.
  3. Ffprobe it is a simple multimedia stream analyzer, that gathers the information from multimedia streams and prints it in a readable way for humans and machines.

  We use FFmpeg to change the videos frame rate and resize the video and remove the duplicated frames from the video to make it easier and reduced the time to detect the anomalies events faster to increase the performance for our model.

- PyTorch is a new deep learning optimized tensor open-source library based on the torch library that is intended for the GPU and the CPU. It's utilized in natural language processing and computer vision. PyTorch provides two unique features:

  1. The first level is similar to NumPy library, the level is the tensor computing with significant acceleration via graphics processing units (GPU), PyTorch is defining a class called Tensor that get by torch. Tensor which uses to store and operate on homogen-

eous multidimensional rectangular arrays of numbers, PyTorch supports various sub-types of Tensors.

2. The second level is Deep neural networks built on a type-based automatic differentiation system that use in Autograd module from PyTorch modules.

- PyTorch Have three modules:

  1. Autograd module PyTorch uses automatic differentiation method, Recorder saves the results of operations and then repeats them backwards to compute the gradients.
  2. Optim module called by torch.optim this module implements different optimization algorithms using for building neural networks, no need to build it from scratch because most of the commonly used methods are already supported.
  3. nn module this module used because though the autograd module can define computational graphs easily and take gradients but this raw can be too low-level for define complex neural network because that the nn module can help instead of autograd module.

- TensorFlow is an open-source machine learning solution with a symbolic math library built on dataflow and differentiable programming that may be used from start to finish. This makes obtaining data, training models, serving forecasts, and refining future outcomes more easier. It can be used for a variety of tasks, although it focuses on deep neural network training. From the advantages of TensorFlow is that it can run on multiple CPUs and GPUs, it is easy deployment of computation across a different platform because its flexible architecture. TensorFlow computations are defined as stateful dataflow graphs that are intended for new approaches of machine learning. The Advantages of TensorFlow is the abstraction so instead of dealing with details for implementing algorithms because that the developer can focus on the general logic, in TensorFlow the TensorBoard visualization suite allows us to inspect and view graphs through an interactive web-based dashboard.

- OpenCV is an open-source computer vision and machine learning toolkit. It was intended to provide a common infrastructure for computer vision applications and to make it easier for commercial products to incorporate machine perception. Companies can easily

use and update the code because OpenCV is a BSD-licensed product. The collection comprises over 2500 optimized algorithms that cover a wide spectrum of both traditional and cutting-edge computer vision and machine learning techniques. These algorithms can be used to detect and recognize faces, identify objects, classify human actions in videos, track camera movements, track moving objects, etc. Scikit-image is an open-source Python image processing package that includes segmentation, geometric transformation, color space manipulation, analysis, filtering, and feature detection methods. It is built to work with Python's NumPy and SciPy numerical and scientific libraries.

- Scikit-learn that called by sklearn that is a free software library for machine learning, it uses a Python consistency interface to deliver a set of fast machine learning and statistical modelling techniques, such as classification, regression, clustering, and dimensionality reduction. It is mostly built in Python, and it heavily relies on NumPy for high-speed linear algebra and array operations and some core algorithms build by Cython to improve the performance.

6.2 Future Frame Prediction

Read Video function, we use the OpenCV2 to get the frames from the video then we can resize it to our desired size and turn it into a grayscale and can control the selected frames. Preprocessing Video function, we use the FFMPEG open-source library mentioned above to be able to remove the still frames and duplicate frames that have no extra meaning to the model, we also use the FFMPEG to change the frame rate of the video to the desired frame rate, and we used it to write a video from a list of frames, and we also use FFMPEG also to change the codec of the video and format of the video and fix the extensions. We also use our extracted videos and turned it into a group of clips, each clip contains 5 frames, from the entire dataset we iterate through each frame and take five frames from the current frame to the frame + 5 we stored all of these clips into a file so we can feed it into the model. We use the TorchVision transform to normalize the videos from -1 to 1 and resize the images and flip the frames horizontally for data augmentation, then we transform them to tensors so the machine learning model can use them. The Data Loader object in the PyTorch library takes the array of clips and returns the clip transformed and preprocessed successfully to the model directly. After a vast research and gathering of resources, most implementations of Generative Adversarial Networks are

done by the PyTorch Library, as all courses have used the PyTorch Library to build Generative Adversarial Networks. We implemented the Laplacian method mentioned above by using the PyTorch Library to determine the edges of the images fed into the network of the machine learning model. Using the skimage library we used the functions PSNR and SSIM that compare between two frames where they produce a value from the two functions that are then used to get the total SSIM and PSNR, we also use the numpy mean in order to calculate the Mean Squared Error where you subtract one frame from the other frame and then you square the value and then divided by the number of frames used which is two. We extract a fake video from the generator to provide the anomaly detection with the fake video.

6.3 Anomaly Detection

The function Get Video Frame takes a video path, it uses the OpenCV2 to split the video to sixteen frame clips. The clips extracted from the function are passed to the C3D Pre-trained model, pre-trained on the Sports-1M, where it has the video features extracted. The features are then passed to the function Interpolate, this function normalizes the input to the same 32x4000 features by using Segmenation32(4096D). The normalized features are passed to the Fully Connected Layer for scoring. The scoring is extracted and passed to the Exterpolate function, this function translates the scores to the timing and produces the video with the graph of anomaly scores the function uses numpy library for switching the scores to a numpy array for showing the results. We implemented the C3D and the Multi-Instance Learning with TensorFlow Library for machine learning. We evaluate the scores and the timing and the correctness of the output by Receiver Operating Characteristics(ROC) Curve We use the roc_auc_score that takes the predicted frames and the label of the frames and produces the false positive rate and the true positive rate. We use the sklearn.metrics.auc that takes the false positive rate and true positive rate and produces the accuracy of the model.

6.4 Web Application

We have implemented a web application that deploys both of our models the future frame prediction model and the anomaly detection model this web application is used to show how our pipeline works in full and also explain how each of the models work individually, We used the flask framework in order to build up this website, we also use the colab platform

in order to deploy the models as our computational power is not sufficient for such models, we used the ngrok to host the web application on the internet through the colab platform, there are three main functions:
The first function takes a video and predict future frames and merge these future frames in order to make a fully predicted video with the frames.
The second function, takes the video and determines the anomaly through the scoring functions on the graph given through the timeline of the video.
And the third function, takes the video and calculates its anomaly score and then predicts the future frames merges them into one video and then calculates anomaly score for that video and then compares the two videos to each other where it shows them on the website side by side as a representation for the pipeline in action.

7 Evaluation

7.1 Generative Adversarial Network Evaluation

The main two types of evaluation methods for images are quantitative and qualitative evaluation. Qualitative evaluation is functional evaluation for the applied processing algorithm on the images. So, it interests in evaluating the image quality. The main idea of quantitative evaluation is to compare the differences between the generated image and the ground truth image, but cannot describe the visual quality (human vision) of the image. In contrast, Qualitative evaluation is a visual examination of images depending on human vision. This evaluation method suffers from some drawbacks as the evaluation of images depending on human vision is expensive, and the evaluation may be biased towards the overfit models. According to this, quantitative evaluation is more accurate than qualitative evaluation. So, we used the quantitative evaluation in this paper using three evaluation metrics: structural similarity square error (SSIM), peak signal to noise ratio (PNSR), and mean squared error (MSE).

7.2 Mean squared error (MSE)

It is the most traditional and simple method for image quality evaluation. The main idea is to compute the squared difference between the original image and ground truth comparing pixel by pixel, then calculating the average. The lower values closer to zero are better for MSE.

$$\mathbf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

All images with the same MSE does not mean that all images contain the same noise and distortions. So, MSE does not consider the structural similarity of the generated images and poorly correlated to visual perception.

7.3 Structural Similarity Index Measure

Structural Similarity Index Measure(SSIM) is a perception-based evaluation metrics considering any change in the perception in structural similarity in images as a distortion. The main purpose of SSIM is to extract structural information from images. SSIM gets information parameters

from the image as luminance contrast, and structure. Luminance is parameter shows that distortion of the generated image is less visible in the edges. contrast is parameter shows that distortion of the generated image is less visible in the texture. Finally, the structure is computed using the other parameters. The higher the SSIM is the better in the quality of the generated image.

7.4 Peak Signal to Noise Ratio

Peak Signal to Noise Ratio(PNSR) is an evaluation metrics uses MSE to calculate the ratio between the maximum pixel value and the maximum noise of the generated image. Therefore, the maximum pixel value considered as the original data and the maximum noise generated considered as the distortion in the generated image. So, PNSR is used to measure the reconstruction quality of the generated image. The higher the PNSR is the better in the quality of the generated image. The equation showed below where m is the maximum value (peak value).

$$\text{PSNR}=10\ log_{10}\left(\frac{m^2}{\text{MSE}}\right)\ \ db.$$

7.5 Anomaly Evaluation

Considering the current models on the anomaly detections, the common accuracy methods used are the direct calculations of the accuracy through the TP, FP, TN and FN, but these methods are not possible on our method the Multi-Instance Learning as we need to find the anomaly temporal wise.

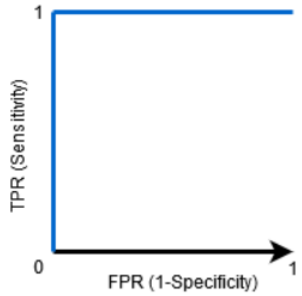
7.6 Receiver operating characteristic (ROC) Curve

The ROC curve is abinary classification issue evaluation metric. It is a probability curve that displays the TPR against the FPR at different threshold values, effectively separating the signal from the noise. The Area Under the Curve (AUC) is a summary of the ROC curve that measures a classifier's understanding of the difference between classes.

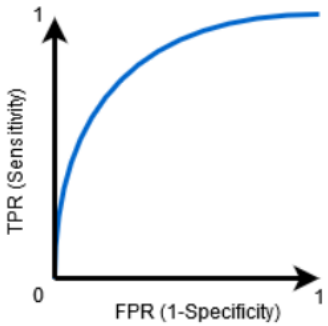
> *The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.*

The ROC Curve specifies the goodness of the model where if the curve approaches the upper left corner, then the ROC Curve is approaching the

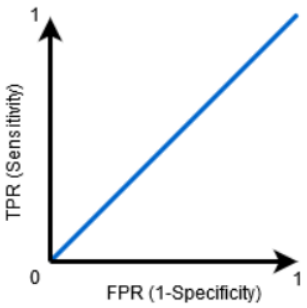
best possible result. When the AUC is equal to one, the classifier is able to differentiate between every positive and negative video in the dataset.



If the ROC Curve is approaching the upper left corner but not perpendicular to the X-axis then it is good but not as good as when it is completely parallel to the Axes, when the AUC is more than the 0.5 threshold and less than the one threshold then this classifier will have a higher chance to classify the negative and positive videos in the dataset.



If the ROC Curve is approaching the middle of the axes then this model is not efficient to classify anything as it will give random results and not accurate at all, so when the AUC is equal to the 0.5 threshold then this classifier will never be able to differentiate between either the positive or the negative videos in the dataset which renders the model useless and not efficient for any classifying



7.7 Our Results

The results were split into three sections for clarification, as our model was split into two separate models and then integrated into a single framework.

7.7.1 Future Frame Results

. Considering the current runs that have been done by the current resources from Google Colaboratory, we worked on future frame prediction, which is the first phase of our pipeline, and trained it on different datasets with different FPS methods. Our results are shown in Table 1.

Table 2: Results on Future Frame Predictions

Dataset	*FPS*	*Epochs*	*SSIM*	*PSNR*	*MSE*
Chunk Avenue	5	21	0.9424	29.40	1.41
Chunk Avenue	10	50	0.9715	35.47	0.44
ShanghaiTech	3	259	0.9528	31.65	1.02
ShanghaiTech	10	10	0.9595	30.23	1.07

7.7.2 Anomaly Detection Results

We tested the model on the Avenue Dataset and achieved an accuracy of 60.47% (Fig. 3); for the ShanghaiTech Dataset, we achieved an accuracy of 88.27% (Fig. 3). Results are shown on the ROC curves.

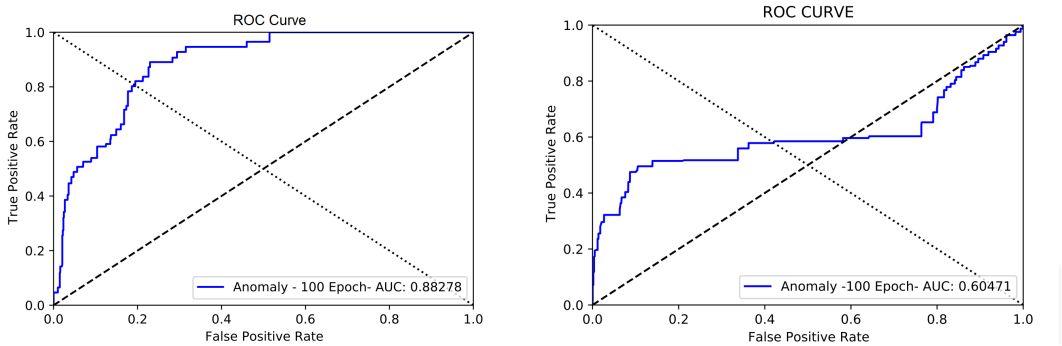


Figure 6: ROC curves on anomaly detection: The left curve corresponds to the ShanghaiTech Dataset; the right curve corresponds to the Avenue Dataset

7.7.3 Proposed Integration Model Results

We used the trained generator from the retrospective cycle GAN; for every four ground-truth frames, three frames were generated and then combined to form a video that contains the generated frames with some ground-truth frames. The clips produced from this were used by the anomaly model to produce an ROC curve (Fig. 4), showing its accuracy. In the Avenue Dataset, the accuracy improved by 8.471% to 68.94% compared with the anomaly detection model with no future frame prediction, while the ShanghaiTech Dataset showed no noticeable improvement at 88.26%. We proved that we can depend on the future frames generated from the GAN model for anomaly prediction.

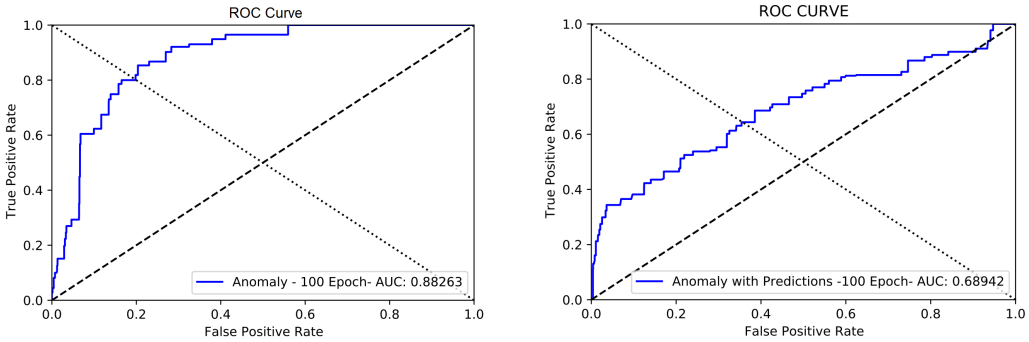


Figure 7: ROC curves on anomaly prediction: The left curve corresponds to the ShanghaiTech Dataset; the right curve corresponds to the Avenue Dataset

8 Experiments

8.1 Experiment 1: 3D Convolutional Neural Network Classification

Video Analysis is a general technique used to extract information from a video or a group of photos, this technique is split into two specific types which are the Video Content Analysis and the Video Motion Analysis. Video content analysis also known as video analysis or video analytics, is the capability of automatically analyzing video to detect and determine temporal and spatial events. Video motion analysis is a technique used to get information about moving objects from video. Examples of this include gait analysis, sport replays, speed and acceleration calculations and, in the case of team or individual sports, task performance analysis. The motion analysis technique usually involves a high-speed camera and a computer that has software allowing frame-by-frame playback of the video. In this experiment we start our work with the introduction to video analysis where we attempt to learn how can we process the images and process them using the pre-processing and classification techniques in order to classify such videos towards their correct label. The dataset used in this experiment is the UCF-101 Dataset, UCF101 is an action recognition data set of realistic action videos, collected from YouTube, having 101 action categories. This data set is an extension of UCF50 data set which has 50 action categories. With 13320 videos from 101 action categories, UCF101 gives the largest diversity in terms of actions and with the presence of large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, illumination conditions, etc, it is the most challenging data set to date. As most of the available action recognition data sets are not realistic and are staged by actors, UCF101 aims to encourage further research into action recognition by learning and exploring new realistic action categories.

Figure 8: UCF Action Recognition Dataset Classes

We used only certain classes from the entire dataset for the reason of computational power, we use the categories that are Boxing Speed Bag, Playing Guitar, Tai Chi, Jump Rope ,Mopping Floor, Knitting, Horse Race, Skydiving, Handstand Pushups, Diving. The pre-processing technique used in this experiment: Resize each video frame to 128*128 pixels. We use a preprocessing technique at which it summarizes the video to a number of frames which is calculated by the following formula:

```
[x * TotalFrames / SummarizedFrames(15) for x in range(15)]
```

This formula will give you the numbers of the frames that you need to extract from the video, which is then used to extract these frames for usage for the classification process. 3D Convolutional Neural Network features for every 15-frame video clip followed by l2 normalization. To obtain features for a video segment, we take the average of all 15-frame clip features within that segment. To calculate the low-level feature representations, 3D convolutions apply a 3-dimensional filter to the dataset, which moves in three directions (x, y, and z). A 3-dimensional volume space, such as a cube or cuboid, is the output shape. They are useful for detecting events in videos, 3D medical pictures, and so on. They are not just for 3d space; they may also be used with 2d space inputs like photos.

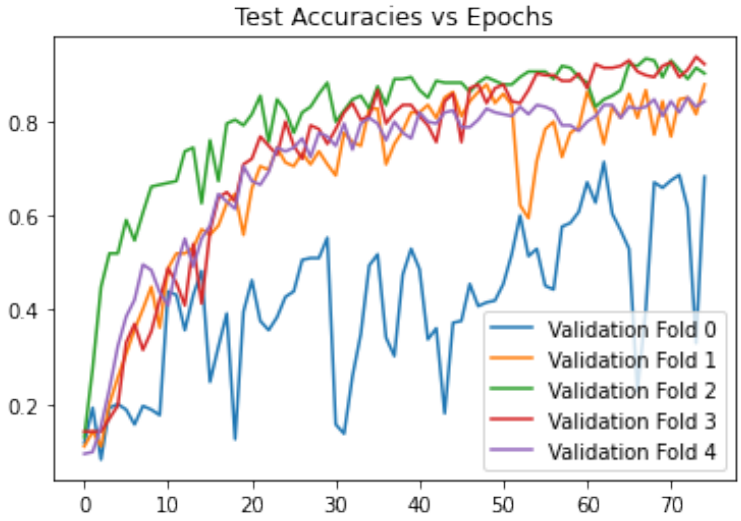
8.1.1 Model Summary

```
Model: "sequential_20"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv3d_62 (Conv3D)           (None, 32, 32, 15, 32)    2624
_________________________________________________________________
activation_48 (Activation)   (None, 32, 32, 15, 32)    0
_________________________________________________________________
conv3d_63 (Conv3D)           (None, 32, 32, 15, 32)    27680
_________________________________________________________________
activation_49 (Activation)   (None, 32, 32, 15, 32)    0
_________________________________________________________________
max_pooling3d_36 (MaxPooling (None, 10, 10, 5, 32)     0
_________________________________________________________________
dropout_36 (Dropout)         (None, 10, 10, 5, 32)     0
_________________________________________________________________
conv3d_64 (Conv3D)           (None, 10, 10, 5, 64)     55360
_________________________________________________________________
activation_50 (Activation)   (None, 10, 10, 5, 64)     0
_________________________________________________________________
conv3d_65 (Conv3D)           (None, 10, 10, 5, 64)     110656
_________________________________________________________________
activation_51 (Activation)   (None, 10, 10, 5, 64)     0
_________________________________________________________________
max_pooling3d_37 (MaxPooling (None, 3, 3, 1, 64)       0
_________________________________________________________________
dropout_37 (Dropout)         (None, 3, 3, 1, 64)       0
_________________________________________________________________
flatten_18 (Flatten)         (None, 576)               0
_________________________________________________________________
dense_36 (Dense)             (None, 512)               295424
_________________________________________________________________
dropout_38 (Dropout)         (None, 512)               0
_________________________________________________________________
dense_37 (Dense)             (None, 10)                5130
=================================================================
Total params: 496,874
Trainable params: 496,874
Non-trainable params: 0
```

Cross-validation is a resampling technique for evaluating machine learning models on a small sample of data. The process includes only one parameter, k, which specifies the number of groups into which a given data sample should be divided. As a result, the process is frequently referred to as k-fold cross-validation. First, we used to split the train test split in order to produce a 20 percent test dataset and 80 percent train dataset. Now we use the cross-validation in order to reach the best parameters possible and avoid overfitting produced by the traditional models.

8.1.2 Results

We see here the graph of the five folds performed on the 3D Convolutional Neural Networks done on the UCF-101 dataset.



The major advantages of the 3D CNN Models are:

- 3D Classifications (The ability of the model to work on entire videos not frames)
- High accuracy in image recognition problems
- Powerful model for learning representations for volumetric data

The major disadvantages of the 3d CNN Models are:

- Classification of images with different positions
- Adversarial examples
- Coordinate frame
- requires a high computational cost and consumes lots of memory.

Spatial: Spatial refers to space, where the anomaly happens in a certain place in the presented frame.
Temporal: Temporal refers to time, where the anomaly happens in a certain time in the presented video.

8.1.3 Our Resolution from this Experiment

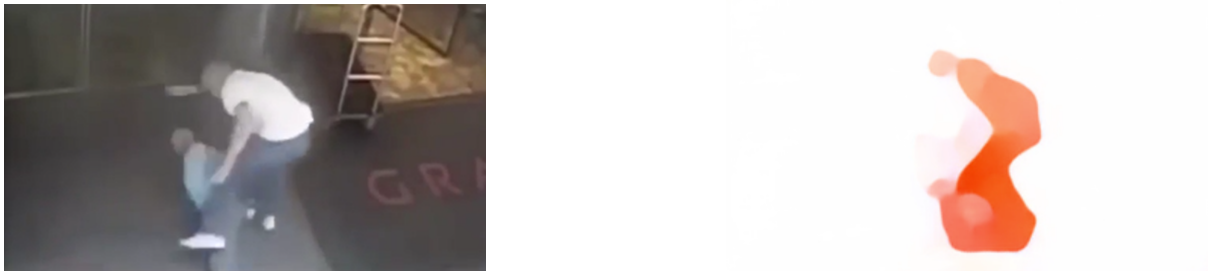
In the UCF-101 dataset it was a dataset that has small videos in a large variety where each category would describe an action, but in the anomaly datasets they are real-world videos that happen in a small time in the video where the temporal labels are not available and it is hard to obtain by hand in a large dataset such as the UCF-101 Crime and the Avenue Dataset, which then drives us to work into two other techniques that are weakly supervised learning and un-supervised learning.

8.2 Experiment 2: Future Frame Prediction for Anomaly Detection - A New Baseline

After a vast research from Experiment 1 results, we needed a model that works with the unsupervised or weakly supervised data, we found the Future Frame Prediction for anomaly detection baseline, this method detects the anomaly by predicting one future frame and compare it with the ground truth of that prediction, if the reconstruction is correct there are no anomalies, if the reconstruction is not correct then there is an anomaly detected.

8.2.1 Optical Flow

We use the FlowNet for optical flow estimation as we produce the optical flow for the frames where the FlowNet renders the background as white and any movement presented in the video is turned into the Optical Flow of the color orange to red according to the rapidity of the movement.



(a) Without Optical Flow. (b) With Optical Flow.

Figure 9: Comparison Between Optical Flow and Real

And to reconstruct the frames we use the pix2pix and U-Net in order to reconstruct the frames where the optical flow frames are constructed successfully to a new future frame that is compared to the ground truth.

This method detected the anomalies only and had to be improved to satisfy the future frame prediction, our purpose was to predict future frames in the first place and then detect the anomalies, and predict more than one frame, which then drove us to research more on the future prediction methods, which is mentioned in the next Experiments.

8.3 Experiment 3: Future Frame Prediction Method

In this experiment we use the Retrospective Cycle GAN method (which is mentioned in detail above). This experiment was the first experiment done using the future frames method, we have not had enough knowledge about the Generative Models, so we tried different approaches on different datasets from the approaches mentioned before, to reach the best result using this method.

8.3.1 Pre-processing

Normalizing images for the input sequence intensities to be [-1,1], flipping the input sequence horizontally with a probability of 0.3 for data augmentation. We resize all the dataset images to 256x256. There are always four frames passed to the Generator in order to produce new frames.

8.3.2 Training Details

We use here the Adam optimizer for the mini-batch stochastic gradient descent method with momentum parameters where the hyperparameters B1 = 0.5 and B2 = 0.999, with a batch size of 3 and learning rate is 0.0003 with linearly decay per every 50 epochs to balance different losses the hyperparameters Lambda1 = 0.005, Lambda2 = 0.003 and Lambda3 = 0003 these parameters' usage is mentioned above in detail, the Leaky ReLU has been set to 0.2 to the negative slope.

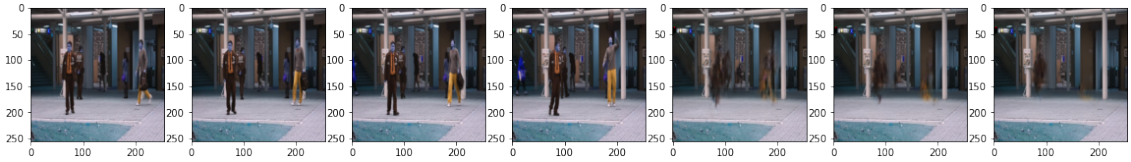
8.3.3 Retrospective Cycle GAN Case Studies

- 1- **UCF Crime Case**: We used the UCF Crime Dataset and worked on 10 classes only due to computational limitations, we followed the same approach as in Experiment 1 pre-processing, where we extract the summarization of the videos in fifteen frames where these frames will describe the video in full. We worked on it for 50 Epochs, after the 50 epochs have concluded we did not obtain satisfying results, as we can see here the produced frame is not clear at all and dim,

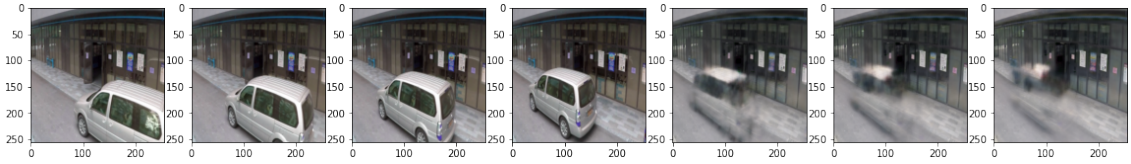
which is not acceptable at all, the photo does not show any information that is available to satisfy classification models.

Conclusion: We faced many problems, one of them is the dataset videos were different in locations as each video occurs in different scenes not in the same scene, which then confuses the Generator as it does not know where it is in the current time, also the temporal annotations were not present and not accurate, also the dataset was really large where we couldn't run it fully on the current computational resources as it required a huge amount of RAM and VRAM, the videos' quality was not high as some videos were in a dim lighting where the generator would produce black videos in the future frame prediction so no action were produced in the future frame which then lowers the accuracy and removes the actions from the video.
The fifteen-frame approach was not accurate and not good with such model, because the videos was really long and the frames were not related enough for the generator, actions might differ in the temporal realm, as the summarized frames were not as close as required.

- 2- **Avenue Case and Shanghai Tech(3fps)**: Avenue Dataset was a dataset for anomaly detection where the dataset was in a very high quality and very clear and has the same scene and the camera was stationary and the dataset itself was lightweight, which solves many problems that was present in the UCF-Crime dataset, we were excited to work on this dataset due to its clearance and its quality. We use another approach in order to solve the problem of the first Case which is the video summarization where the frames were not related to each other good enough, so we took the 25fps and changed the frame rate to 3fps, so we can let the 3 frames produced describe an entire second instead of the 25 fps method where 25 frames were used to describe the entire second. Then we group 4 frames together in order to produce a future frame as mentioned in the pre-processing section in detail above. We trained the model on the 3 fps method for 200 epoch and achieved a Structured Similarity Index Measure 0.93.

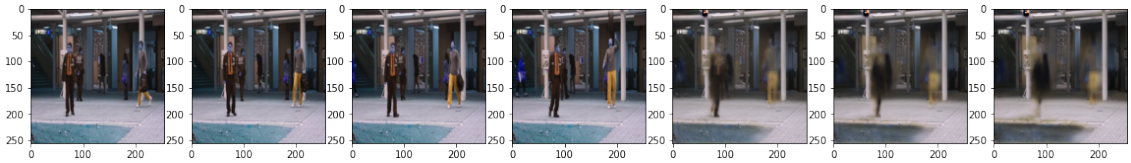


Shanghai Tech Dataset was a dataset for anomaly detection as well where the dataset was always clear and in a very high quality, it consists of 13 scenes where anomalous actions happen, we only work on 2 large scenes of the Shanghai Tech Dataset due to computational limitations, the 3fps method on this dataset produces 0.9528.

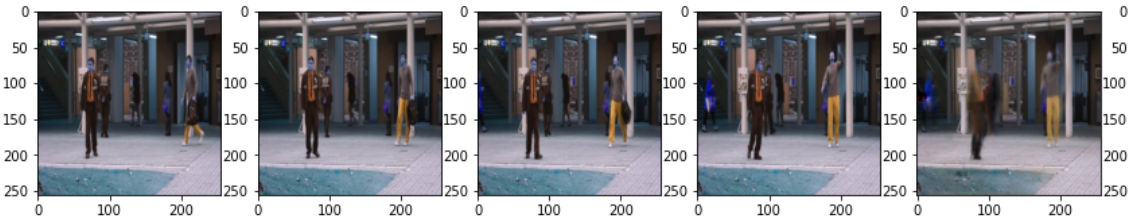


In this method we were able to perform a lot of epochs on the dataset as the dataset was really small and lightweight, the problem was not producing a good frame that we can depend on anomaly-wise, so we looked for other methods for the frame prediction method.

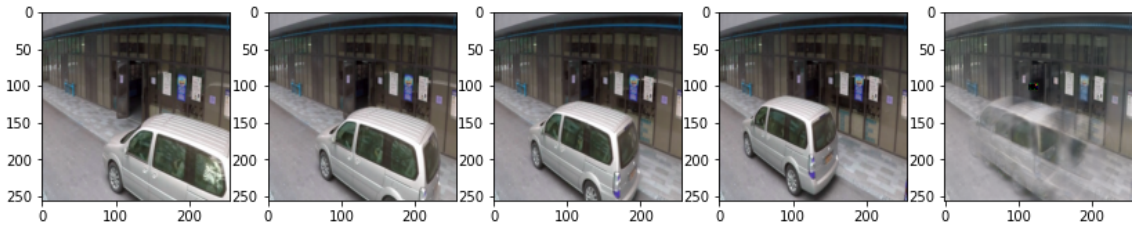
- 3-**Avenue Case (5fps)**: In this trial we use another approach where we use a higher frame rate expecting to get a higher Structured Similarity Index Measure, after the trial of 21 epochs we achieved the Structured Similarity Index Measure of 0.9424.



- 4-**Avenue Case and Shanghai Tech(10fps)**: In this trial we use another approach where we use a higher frame rate expecting to get a higher Structured Similarity Index Measure more than the one in case 3, after the trial and performing the Structured Similarity Index Measure of 0.9715 which is the highest Structured Similarity Index Measure that we have achieved till now on this dataset.



And on the Shanghai Tech Dataset we achieved the Structured Similarity Index Measure of 0.9595



The reason behind using 10fps was because of the time, because as we increase the frame rate the computational power increases as well, while the time also increases while producing frames in more time. Here we reached the best results in the 10fps method as the frames were dependable when predicting anomalies.

8.3.4 Our Resolution from this Experiment

We conclude this long experiment that there is a tradeoff between the time, computational force and the frame rate, this experiment also shows that accuracy also known as Structured Similarity Index Measure goes higher as the framerate increases, the highest Structured Similarity Index Measure was achieved by the 10FPS method on both datasets which are the Shanghai Tech Dataset and the Avenue Dataset. The 3 FPS Method had produced less videos overall in the dataset, where the epoch would go through less videos and the generator would not have trained sufficiently, while the 5 FPS Method produces more videos so the Structured Similarity Index Measure has increased because of the dataset increasing with a vast major amount, while the 10 FPS Method has produced way larger videos which makes the Generative Adversarial Networks have sufficient training data. So, from this experiment we made a thesis that the background would not be mandatory in the current detection datasets, so we try to remove the background and test the current frame prediction method while removing the background.

8.4 Experiment 4: Background Removal

This experiment's purpose is to remove the background and make the model have the full focus on the individuals' movements only while not having to generate the entire background for the scene, focusing on the prediction of the person's movement only. There are a lot of methods that

remove background like the OpenCV but OpenCV did not produce acceptable results, and after more research we came up with the PixelLib that removes the background efficiently as PixelLib uses object segmentation to perform excellent foreground and background separation. It makes it able to edit the background of images and videos using just 5 lines of code that removes the background also the PixelLib is implemented with the Deeplabv3 framework to perform semantic segmentation by using pascalvoc Dataset.

(a) OpenCV (b) PixelLib

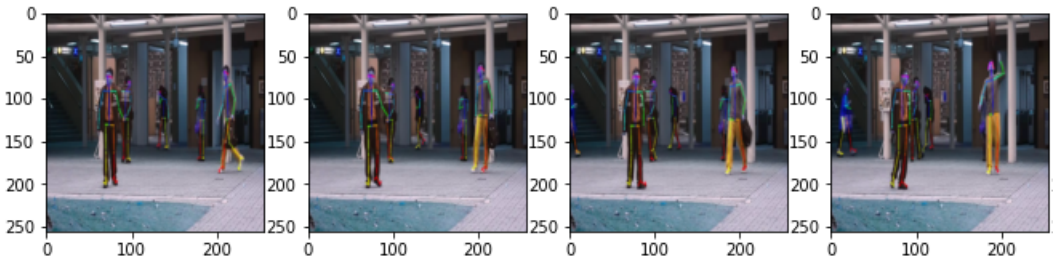
Figure 10: Comparison Between OpenCV and PixelLib

The Advantage of PixelLib is that it does object and person detection first and then removes the background accordingly, which improves the accuracy of the background removal. After removing the background of the entire Avenue Dataset we trained the Generative Adversarial Networks on it, and it did not produce better results. The Background removal have destroyed a couple of equations that are used in the Generative Adversarial Networks Generator, these equations are the Laplacian of Gaussian which detects the edges of the image for the generator and the filter layers from the generator model, which produces future frames that are meaningless and not useful for our cause.

8.5 Experiment 5: Pose Estimation

From the trial on Experiment 4 the background removal produced meaningless frames that are not useful for our cause which is the future frame prediction, it now produces people that are walking aimlessly to the model. However, that drove us to predict only the person's movement with no background, by using the person's pose and classifying it as anomaly or not, so all the persons in the frame will be tracked on his own. Pose estimation is a computer vision technique for predicting and tracking a person's or object's location. This is accomplished by examining a person's or object's stance and orientation in combination. After an exhaustive research, we have reached three different post estimation methods, which

are Detectron2, YOLO3 and YOLO4. The best result was achieved by the Detectron2 pre-trained on the COCO Dataset, which gives the best results in the Detectron2.



We were working at the method of 10FPS and when the pose estimation tracked each person to predict the person's next behavior, after performing the pose estimation the dataset have enlarged way too much which increased the computations required way too much which the computation limitations prevented us from performing it on the 10FPS dataset. When we tried the generation process, there was no change in Structured Similarity Index Measure. To sum up all these experiments up, Experiment 3 is the best approach that we can take and the best method in that approach is the 10FPS method, there are other methods that needs trials, but the computational limitations have prevented us from performing them, these experiments will be mentioned in the future works.

9 Conclusion and Future Work

9.1 Conclusion

In summary, we proposed a framework integrating two models: a future frame prediction method and a weakly supervised deep learning classification method. The framework was used to predict anomalous behavior in surveillance cameras to prevent crime. We used the future frame prediction (Generative Adversarial Network) method to close the gap left by insufficient anomaly prediction methods. By predicting future frames containing anomalous behavior before it happens by a certain duration, we showed the possibility of depending on the predicted frames, as accuracy either increased or did not change using the given datasets.
Our results not only open possibilities in anomaly prevention but also provide a good understanding of what could be done with new improvements in future frame generation methods. In future works, we are looking forward to modifying Generative Adversarial Networks to produce a higher number of frames with higher Structured Similarity Index Measure to save as much time as possible to predict future anomalies much earlier and improve the deep learning method to achieve higher accuracy using the given datasets.

9.2 Future Works

As mentioned in our experiments we have way more possibilities in the future frame prediction methods which is the Retrospective Cycle Generative Adversarial Networks, the possibilities are various and promising, knowing the current results which was run on very low computational force which is GeForce 840M GPU and an i7-4510U CPU, the increase in computational force will achieve way higher results in way less time which will give us the sufficient computations for the current models as these models are power hungry and they need to be contained in order to run on our machines, super computing facilities such as the High Performance Computing Facility in Alexandria Library have not been at all helpful in these experiments as they have locked absolutely everything that can be used in our own benefit, rendering our computations and models that we have ran there totally useless and the results that have been extracted from there were completely broken and meaningless, this drives our future work towards other High Performance Computing Facilities that might give us the required computations for the current model, this is the hardest challenge that we have faced, that is because we haven't reached the end of

the desired number of epochs so that we can have a clear understanding of the complete possibilities that we have reached.
Moving on we have made much more research since then and reached new works that have been published not long ago, which is the Any-Shot Sequential Anomaly Detection in Surveillance Videos paper published at 5 April 2020 which proposes a better anomaly detection model that is way better than the Multi-Instance Learning as this method does not require the entire video to work with the anomalies and detect them, instead it uses the optical flow and YOLO as it performs object detection on the persons and then performs pose estimation to detect the anomalies, this paper will help us to achieve real-time prediction that we originally wanted to do.
On the other hand on the Generative Models, our research show a new paper Long-Term Human Video Generation of Multiple Futures Using Poses that have been released in 1 Jun 2021, this paper predicts the future pose of a human for 4 to 5 whole seconds, this paper show amazing results that we can use in the future in order to increase the timing of the future frames and then predict the anomalies much earlier and prevent them entirely, they present a novel method for generating long-term future videos of multiple futures from an input human video using a hierarchical approach: first predicting future human poses and then generating the future video. They propose a novel network to predict long-term future human pose sequences by using unidimensional convolutional neural network in adversarial training. Also, they propose two additional inputs that allow predicting a variety of multiple futures: a latent code and an attraction point. Finally, videos generated with their predicted poses are also long and multiple. Experimental results on the realism, diversity, and accuracy of the generated poses and videos show the superiority of the proposed method over the state-of-the-art, which then inspires us to work more on that approach where it will be the better model to use when predicting the future and classifying the anomalies afterwards.
We look forward to implement a mobile application that can run the models by optimizing the current models to a point where it is possible to be ran on the mobile devices easily and with no overhead.

References

Aberkane, S. and Elarbi, M. (2019). Deep reinforcement learning for real-world anomaly detection in surveillance videos, *2019 6th International Conference on Image and Signal Processing and their Applications (ISPA)*, IEEE, pp. 1–5.

Adam, A., Rivlin, E., Shimshoni, I. and Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors, *IEEE transactions on pattern analysis and machine intelligence* **30**(3): 555–560.

Adams, A. A. and Ferryman, J. M. (2015). The future of video analytics for surveillance and its ethical implications, *Security Journal* **28**(3): 272–289.

Andrews, S., Tsochantaridis, I. and Hofmann, T. (2002). Support vector machines for multiple-instance learning., *NIPS*, Vol. 2, p. 7.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J. (2016). Netvlad: Cnn architecture for weakly supervised place recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307.

Ashby, M. P. (2017). The value of cctv surveillance cameras as an investigative tool: An empirical analysis, *European Journal on Criminal Policy and Research* **23**(3): 441–459.

Aslan, S., Güdükbay, U., Töreyin, B. U. and Çetin, A. E. (2019). Deep convolutional generative adversarial networks based flame detection in video, *arXiv preprint arXiv:1902.01824* .

Baldwin, D. A. and Baird, J. A. (2001). Discerning intentions in dynamic human action, *Trends in cognitive sciences* **5**(4): 171–178.

Barrett, H. C., Todd, P. M., Miller, G. F. and Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study, *Evolution and Human Behavior* **26**(4): 313–331.

Bulwa, D. (2007). Is it worth the cost?
URL: *https://www.sfgate.com/news/article/Is-it-worth-the-cost-2546948.php*

Byeon, W., Wang, Q., Srivastava, R. K. and Koumoutsakos, P. (2018). Contextvp: Fully context-aware video prediction, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 753–769.

Chaabane, M., Trabelsi, A., Blanchard, N. and Beveridge, R. (2020a). Looking ahead: Anticipating pedestrians crossing with future frames prediction, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2297–2306.

Chaabane, M., Trabelsi, A., Blanchard, N. and Beveridge, R. (2020b). Looking ahead: Anticipating pedestrians crossing with future frames prediction, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Chan, F.-H., Chen, Y.-T., Xiang, Y. and Sun, M. (2016). Anticipating accidents in dashcam videos, *Asian Conference on Computer Vision*, Springer, pp. 136–153.

Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection: A survey, *ACM computing surveys (CSUR)* **41**(3): 1–58.

Coetzer, B., van der Merwe, J. and Josephs, B. (2011). Information management and video analytics: The future of intelligent video surveillance, *Video Surveillance* p. 1.

Cong, Y., Yuan, J. and Liu, J. (2011). Sparse reconstruction cost for abnormal event detection, *CVPR 2011*, IEEE, pp. 3449–3456.

Cui, X., Liu, Q., Gao, M. and Metaxas, D. N. (2011). Abnormal detection using interaction energy potentials, *CVPR 2011*, IEEE, pp. 3161–3167.

Datta, A., Shah, M. and Lobo, N. D. V. (2002). Person-on-person violence detection in video data, *Object recognition supported by user interaction for service robots*, Vol. 1, IEEE, pp. 433–438.

Davenport, J. (2007). Tens of thousands of cctv cameras, yet 80% of crime unsolved.
URL: *https://www.standard.co.uk/hp/front/tens-of-thousands-of-cctv-cameras-yet-80-of-crime-unsolved-6684359.html*

Ding, S., Lin, L., Wang, G. and Chao, H. (2015). Deep feature learning with relative distance comparison for person re-identification, *Pattern Recognition* **48**(10): 2993–3003.

Dowling, C., Morgan, A., Gannoni, A. and Jorna, P. (2019). How do police use cctv footage in criminal investigations?, *Trends and issues in crime and criminal justice* (575): 1–15.

Edwards, R. (2008). Police say cctv is an 'utter fasco'.
URL: *https://www.telegraph.co.uk/news/uknews/1932769/Police-say-CCTV-is-utter-fiasco-as-most-footage-is-unusable.html*

Faber, L. G., Maurits, N. M. and Lorist, M. M. (2012). Mental fatigue affects visual selective attention, *PloS one* **7**(10): e48073.

Fang, J., Yan, D., Qiao, J. and Xue, J. (2019). Dada: A large-scale benchmark and model for driver attention prediction in accidental scenarios, *arXiv preprint arXiv:1912.12148* .

Finn, C., Goodfellow, I. and Levine, S. (2016). Unsupervised learning for physical interaction through video prediction, *arXiv preprint arXiv:1605.07157* .

Fushishita, N., Tejero-de Pablos, A., Mukuta, Y. and Harada, T. (2020). Long-term human video generation of multiple futures using poses, *European Conference on Computer Vision*, Springer, pp. 596–612.

Gill, M. (1994). *Crime at work*, Springer.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S. and Hengel, A. v. d. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714.

Gong, S., Loy, C. C. and Xiang, T. (2011). Security and surveillance, *Visual analysis of humans*, Springer, pp. 455–472.

Goold, B. J. (2004). *CCTV and policing: Public area surveillance and police practices in Britain*, Oxford University Press on Demand.

Gowsikhaa, D., Abirami, S. and Baskaran, R. (2014). Automated human behavior analysis from surveillance videos: a survey, *Artificial Intelligence Review* **42**(4): 747–765.

Gujjar, P. and Vaughan, R. (2019). Classifying pedestrian actions in advance using predicted video of urban driving scenes, *2019 International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 2097–2103.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K. and Davis, L. S. (2016). Learning temporal regularity in video sequences, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742.

Hesse, L. (2002). The transition from video motion detection to intelligent scene discrimination and target tracking in automated video surveillance systems, *Security Journal* **15**(2): 69–78.

Hier, S. P., Greenberg, J., Walby, K. and Lett, D. (2007). Media, communication and the establishment of public camera surveillance programmes in canada, *Media, Culture & Society* **29**(5): 727–751.

Honovich, J. (2019). Is public cctv efective?
URL: *https://ipvm.com/reports/is-public-cctv-efective*

Ionescu, R. T., Khan, F. S., Georgescu, M.-I. and Shao, L. (2019). Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7842–7851.

Jia, X., De Brabandere, B., Tuytelaars, T. and Van Gool, L. (2016). Dynamic filter networks, *NIPS*.

Joachims, T. (2002). Optimizing search engines using clickthrough data, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142.

Kalchbrenner, N., Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A. and Kavukcuoglu, K. (2017). Video pixel networks, *International Conference on Machine Learning*, PMLR, pp. 1771–1779.

Kaur, J. and Das, S. (2020). Future frame prediction of a video sequence, *arXiv preprint arXiv:2009.01689* .

Keval, H. and Sasse, M. A. (2010). "not the usual suspects": a study of factors reducing the effectiveness of cctv, *Security Journal* **23**(2): 134–154.

Kim, J. and Grauman, K. (2009). Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates, *2009 IEEE conference on computer vision and pattern recognition*, IEEE, pp. 2921–2928.

Kwon, Y.-H. and Park, M.-G. (2019). Predicting future frames using retrospective cycle gan, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

La Vigne, N. G., Lowry, S. S., Markman, J. A. and Dwyer, A. M. (2011). Evaluating the use of public surveillance cameras for crime control and prevention, *Washington, DC: US Department of Justice, Office of Community Oriented Policing Services. Urban Institute, Justice Policy Center* .

Lai, Y., Liu, R. and Han, Y. (2020). Video anomaly detection via predictive autoencoder with gradient-based attention, *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp. 1–6.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690.

Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C. and Levine, S. (2018). Stochastic adversarial video prediction, *arXiv preprint arXiv:1804.01523* .

Liang, X., Lee, L., Dai, W. and Xing, E. P. (2017). Dual motion gan for future-flow embedded video prediction, *proceedings of the IEEE international conference on computer vision*, pp. 1744–1752.

Liu, W., Luo, W., Lian, D. and Gao, S. (2018). Future frame prediction for anomaly detection–a new baseline, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6536–6545.

Long, J., Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440.

Lotter, W., Kreiman, G. and Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning, *arXiv preprint arXiv:1605.08104* .

Luo, W., Liu, W. and Gao, S. (2017a). Remembering history with convolutional lstm for anomaly detection, *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp. 439–444.

Luo, W., Liu, W. and Gao, S. (2017b). A revisit of sparse coding based anomaly detection in stacked rnn framework, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 341–349.

Mathieu, M., Couprie, C. and LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error, *arXiv preprint arXiv:1511.05440* .

Morgan, A. and Coughlan, M. (2018). Police use of cctv on the rail network, *Trends and Issues in Crime and Criminal Justice* **1**(561): 1–18.

Naphade, M., Tang, Z., Chang, M.-C., Anastasiu, D. C., Sharma, A., Chellappa, R., Wang, S., Chakraborty, P., Huang, T., Hwang, J.-N. et al. (2019). The 2019 ai city challenge., *CVPR Workshops*, Vol. 8.

Näsholm, E., Rohlfing, S. and Sauer, J. D. (2014). Pirate stealth or inattentional blindness? the effects of target relevance and sustained attention on security monitoring for experienced and naïve operators, *PLoS One* **9**(1): e86157.

of Minnesota, U. (2019). Umn dataset.
URL: *http://mha.cs.umn.edu/*

Patraucean, V., Handa, A. and Cipolla, R. (2015). Spatio-temporal video autoencoder with differentiable memory, *arXiv preprint arXiv:1511.06309* .

Piza, E. L., Welsh, B. C., Farrington, D. P. and Thomas, A. L. (2019). Cctv surveillance for crime prevention: A 40-year systematic review with meta-analysis, *Criminology & Public Policy* **18**(1): 135–159.

Prenzler, T. and Wilson, E. (2019). The ipswich (queensland) safe city program: an evaluation, *Security Journal* **32**(2): 137–152.

Qiang, Y., Fei, S., Jiao, Y. and Li, L. (2020). Anomaly detection of predicted frames based on u-net feature vector reconstruction, *Journal of Physics: Conference Series*, Vol. 1627, IOP Publishing, p. 012014.

Ratcliffe, J. H., Taniguchi, T. and Taylor, R. B. (2009). The crime reduction effects of public cctv cameras: a multi-method spatial approach, *Justice Quarterly* **26**(4): 746–770.

Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C. and Sebe, N. (2017). Abnormal event detection in videos using generative adversarial nets, *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 1577–1581.

Sagar, A. (n.d.). High resolution video generation using spatio-temporal gan.

Sandhu, A. (2019). 'i'm glad that was on camera': a case study of police officers' perceptions of cameras, *Policing and society* **29**(2): 223–235.

Sasse, M. A. (2010). Not seeing the crime for the cameras?, *Communications of the ACM* **53**(2): 22–25.

Scheitle, C. P. and Halligan, C. (2018). Explaining the adoption of security measures by places of worship: Perceived risk of victimization and organizational structure, *Security Journal* **31**(3): 685–707.

Shah, A. P., Lamare, J.-B., Nguyen-Anh, T. and Hauptmann, A. (2018). Cadp: A novel dataset for cctv traffic camera based accident analysis, *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, pp. 1–9.

Sultani, W., Chen, C. and Shah, M. (2019). Real-world anomaly detection in surveillance videos.

Surette, R. (2014). *Media, crime, and criminal justice*, Cengage Learning.

Thomas, J. J. and Cook, K. A. (2006). A visual analytics agenda, *IEEE computer graphics and applications* **26**(1): 10–13.

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks, *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Tulyakov, S., Liu, M.-Y., Yang, X. and Kautz, J. (2018). Mocogan: Decomposing motion and content for video generation, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535.

Villegas, R., Yang, J., Hong, S., Lin, X. and Lee, H. (2017). Decomposing motion and content for natural video sequence prediction, *arXiv preprint arXiv:1706.08033* .

Vondrick, C., Pirsiavash, H. and Torralba, A. (2016). Generating videos with scene dynamics, *arXiv preprint arXiv:1609.02612* .

Wang, D., Yuan, Y. and Wang, Q. (2019). Early action prediction with generative adversarial networks, *IEEE Access* **7**: 35795–35804.

Wang, Y., Jiang, L., Yang, M.-H., Li, L.-J., Long, M. and Fei-Fei, L. (2018). Eidetic 3d lstm: A model for video prediction and beyond, *International conference on learning representations*.

Welsh, B. C., Farrington, D. P. and Taheri, S. A. (2015). Effectiveness and social costs of public area surveillance for crime prevention, *Annual Review of Law and Social Science* **11**: 111–130.

Yao, Y., Wang, X., Xu, M., Pu, Z., Atkins, E. and Crandall, D. (2020). When, where, and what? a new dataset for anomaly detection in driving videos, *arXiv preprint arXiv:2004.03044* .

Yao, Y., Xu, M., Wang, Y., Crandall, D. J. and Atkins, E. M. (2019). Unsupervised traffic accident detection in first-person videos, *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 273–280.

Ye, M., Peng, X., Gan, W., Wu, W. and Qiao, Y. (2019). Anopcn: Video anomaly detection via deep predictive coding network, *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1805–1813.

10 Appendix

in the International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) Chairs are Ayman Bahaa-Eldin, Ashraf AbdelRaouf, Nada Shorim, Randa Osama and Shereen Essam Elbohy at the 26-27 May 2021, We have published a paper named Early Anomaly Prediction in Surveillance Cameras for Security Applications which is the same title as our graduation project, you can check it below.

Early-Anomaly Prediction in Surveillance Cameras for Security Applications

1st Mario Emad
Department of Computer Science
Helwan University
Cairo, Egypt
mario_20170396@fci.helwan.edu.eg

2nd Michael Ishack
Department of Computer Science
Helwan University
Cairo, Egypt
mikeal_20170399@fci.helwan.edu.eg

3rd Mohamed Ahmed
Department of Computer Science
Helwan University
Cairo, Egypt
mohamed_20170412@fci.helwan.edu.eg

4th Mohamed Osama
Department of Computer Science
Helwan University
Cairo, Egypt
mohamed_20170414@fci.helwan.edu.eg

5th Mohamed Salah
Department of Computer Science
Helwan University
Cairo, Egypt
mohamed_20170446@fci.helwan.edu.eg

6th Ghada Khoriba
Department of Computer Science
Helwan University
Cairo, Egypt
ghada_khoriba@fci.helwan.edu.eg

***Abstract*—In the last decade, the number of surveillance cameras has increased significantly, with much research conducted to automate the process of surveillance, as humans cannot manage to monitor all these cameras individually, which may cause errors in public safety or abnormal situations. Also, humans may overlook key details in such abnormal behaviours in surveillance cameras. The proposed approach predicts abnormal behaviour using generative adversarial networks (GANs). GANs are trained using different datasets that contain various behaviours to predict future frames. These future frames are transmitted to a deep learning neural network to classify them as normal or abnormal activities, and future anomalies can be detected before they happen. Our initial results show that depending on the future frames extracted by the GAN model is possible, as these extracted frames either improve the accuracy of the detection model or do not affect it, but they can also be further enhanced to detect more frames at a longer duration and predict anomalies before they happen. Anomalies in surveillance will not only be detected but also predicted before they happen, which will result in the prevention of crimes, reductions in surveillance costs and a safer environment overall.**

***Keywords*—Abnormal Behaviour, Computer Vision, Anomaly Detection, Criminal, Future Frame Prediction, Security, Surveillance, Video Analysis**

I. Introduction

Security is one of the most important concerns worldwide, which includes reducing the number of victims and protecting vital places, as the crime rate has increased in the past three years [1]. Thus, it is a high-profile problem from a social and industrial perspective; the demand for surveillance systems is high. These surveillance systems need to be supported by predictive analysis and the detection of criminal or suspicious activities. The availability of higher-quality but affordable cameras and surveillance systems has recently increased. These cameras have shifted from human-based activities to application-based cameras such as motion detection, facial recognition, and other technologies. However, concerns regarding surveillance systems have increased because of crimes

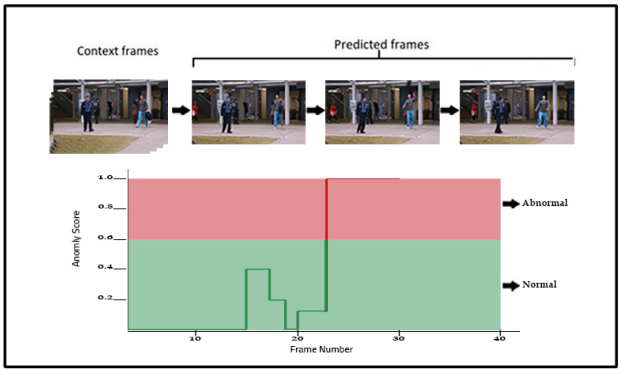


Fig. 1. Proposed Model: Detection

that occur daily, and while cameras have already been installed in many locations, their effectiveness is not sufficient for the current demand by authorities, as most crimes either go unnoticed or are neglected by surveillance security guards. Recent research (mentioned in the next section) found that public cameras do decrease crime but not at an acceptable rate, which has driven researchers to innovate new machine learning techniques to reduce crime or detect it. Recent studies have focused on two approaches. The first refers to the possibility of detecting anomalies by applying the proposed models on existing surveillance videos; the second pertains to the possibility of predicting future frames and comparing them to the ground truth to detect anomalies in real time. The outcomes of these approaches bring us to Honovich [7], who proposed that closed-circuit television (CCTV) surveillance systems are rather used to solve crimes than detect them. Honovich also said that camera network operators should focus on solving rather than preventing crime, so instead of catching criminals in the act, camera networks are used to answer the questions 'What happened?' and 'Who are involved?' Because

978-1-6654-1243-8/21/$31.00 © 2021 IEEE

of this, researchers have been working on preventing crimes by implementing systems that detect crimes after they happen, which is not relevant to the existing problem. This drives the challenge of the research to another level of fighting crime. This study proposes a model to enhance anomaly detection by predicting future crimes. The pipeline shows the details of each step undergone by the proposed architecture, starting from the collection of context frames. The frames are fed into the adversarial generator, which produces future predicted frames. These frames are then merged with past frames and fed into the C3D Feature Extractor for feature extraction. The features are then passed to a fully connected neural network to produce an anomaly score that is used to discover anomalies (Fig. 2).

II. Related Work

We have split the related work into two subsections, as our research depends on two main methods: future frame prediction and anomaly detection. The first method is used to predict future frames using a machine learning model mainly involving auto-encoders, GANs and deep neural networks. The second method is used to detect anomalies in the videos and assign them to a class of the corresponding anomaly. Both models are integrated into a single framework in this research.

A. Future Frame Prediction

Lee et al. [14] proposed the stochastic adversarial video prediction (SAVP) architecture for video prediction, which combines GANs and variational auto-encoders (VAEs). GANs are trained with randomly drawn input whereas VAEs only observe ground-truth images. Kaur et al. [9] used SAVP as the baseline architecture, which is modified so that it could predict the future for more time stamps and also addressed the mode collapse problem. The architecture had two manifold guided generative adversarial networks (MGGANs) and a VAE. The decoder of the VAE was also a generator. An MGGAN integrates a guidance network into the existing GAN architecture. Lotter et al. [16] proposed a method of deep convolutional recurrent neural network inspired by the principles of predictive coding. The model was trained to predict the next-frame video prediction with the belief that prediction was an effective objective for unsupervised learning. Liu et al. [14] and Ganokratanaa et al. [5] both used optical flow and GAN to predict future frames and compare the ground-truth optical flow with the predicted one, so any unknown event is considered an anomaly. Both models are effective in real-time surveillance videos. Chaabane et al. and Gujjar et al. [2], [6] worked on predicting whether pedestrians will cross in front of a vehicle. The authors proposed a multitask model consisting of two stages. The first is passing on the video to an encoder/decoder network for preprocessing and then predicting the next frames. The second one is using the predicted frames to classify the pedestrians' future actions. Landi et al. [13] explored the impact of considering spatiotemporal tubes instead of whole-frame segments for detecting anomalous behaviours in surveillance cameras. The authors used bounding box supervision in both training and test sets. Their experiments showed that a network trained with spatiotemporal tubes performs better than one trained with whole-frame videos. Roy et al. [22] proposed a GAN approach to predict trajectories of vehicles at both signalised and nonsignalised intersections. This approach produced the most acceptable future trajectory among past states of the art in predicting trajectory. Lu et al. [18] proposed a novel sequential generative model based on VAE for future frame prediction with convolutional long short-term memory(LSTM). This was the first work that considered temporal information in a future frame prediction–based anomaly detection framework Lai et al. [12] presented a new framework to detect anomalies in surveillance videos. The authors proposed a new two-branch predictive autoencoder, which included a reconstruction decoder and a prediction decoder, to generate future frames and carry out anomaly detection. This model unified reconstruction and prediction models in an end-to-end framework. Kwon et al. [11] proposed an architecture using two discriminators with one generator, working with forward and backward propagated methods, rendering the model with four discriminators and two generators working in parallel to generate a future frame. The produced frames' structured similarity is higher than that of state-of-the-art methods. We used this method with modifications as described in section 3.

B. Anomaly Detection

Chong et al. [4] built on the convolutional auto-encoder architecture by saving the temporal ordering of frames using convolutions and modelling the temporal information at the bottleneck layer with specialised convolutional LSTM layers. Luo et al. [19] proposed a temporal-coherent sparse coding (TCS) framework that saved similarities between frames in videos. They used a special type of RNN (stacked RNN) to avoid nontrivial parameter selection and reduced the computational cost of reconstruction coefficients in the testing phase. Morais et al. [20] built an approach based on leveraging human skeleton trajectories to calculate body movements and body postures and modelled a subprocess using the message-passing encoder-decoder recurrent network(MPED-RNN) model, composed of two RNNs with gated recurrent units(GRU) architecture. Rodriguez et al. [21] proposed a multiscale hierarchical framework to develop understanding at different timescales to improve the performance of future detection; similar to [11], they proposed a model that reproduces the past. They used human post-trajectory as the input, as it captures human movements well enough for this model, which consists of two models that make past and future predictions. At a particular timescale, we combined the predictions from both models to generate a prediction at every time instance. To generate future predictions at timescale 1 (in our setup, timescale 1 represents a time duration of three steps), the model first splits the sequence into smaller subsequences (of length 3) and then makes future predictions (for the next three steps) for these subsequences. These predictions are combined to obtain the future prediction for the complete input sequence

at this timescale. Sultani et al. [23] introduced a weakly supervised approach that created normal and anomalous videos as bags and video segments as instances in multiple-instance learning (MIL) and then extracted features from each bag by 3D convolutions and trained a fully connected network through which they mapped features to anomaly scores. They introduced a ranking loss to make a distinction between anomalous instances. We used this model as our detection method with some changes explained in section 3.

III. Proposed Method

A few recent studies have mentioned the possibility of predicting actions that are future frame–dependent in other fields, such as autonomous driving and action recognition [2], [6], which drives our work to predict anomalous behaviours before they happen. The proposed method focuses on integrating two main methods, which include the future frame prediction method called retrospective cycle GAN [11] and the weakly supervised detection method MIL [23]. They are both implemented in a single framework installed on surveillance cameras to predict future anomalies depending on the future frame extracted from the GAN model. Each frame undergoes some preprocessing to be compatible with the retrospective cycle GAN and is then switched into a sequence of frames, which is taken as an input to the retrospective cycle GAN to predict future frames. These frames then undergo feature extraction using a pretrained feature extraction model; the features extracted are then fed into the MIL model, which then produces a ranking score that is directly proportional to the anomaly. Hence, if the anomaly's value in a frame is high, the ranking score is high, according to a threshold that corresponds to the data's nature.

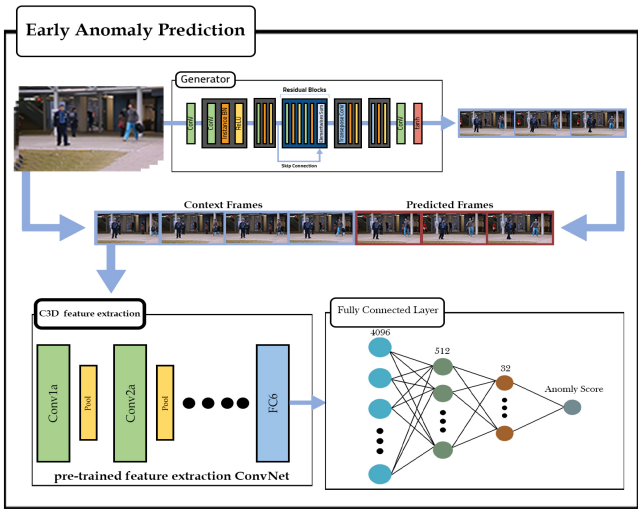


Fig. 2. Pipeline Architecture:predicting future frames passed to deep learning for classification.

A. Future Frame Prediction

We used some methods (see section IV.B) to remove the duplicate frames and reduce the time elapsed by the video. We also adjusted frames per second to summarise the videos and reduced the time elapsed by the model as well as created a list of clips. Each clip consists of certain effective frames. We normalised the intensities of all input sequences N from -1 to 1 and flipped the input sequence frames horizontally with a probability of 30% for data augmentation. We used the Avenue Dataset [17] and the ShanghaiTech Dataset [15].

The retrospective cycle GAN model consists of a generator and two discriminators. The generator takes four types of input in the example: the first type is four real frames, the second type is four real frames reversed (the first frame is last frame, etc.), the third type is three frames and one fake frame, and the fourth type is three frames and one fake frame reversed. These types are all processed inside the generator, which would give the best result possible. We have two types of discriminators: The first type is the frame discriminator, which takes a frame as input and determines whether this frame is fake. The second type is the sequence discriminator, which takes a sequence of frames as input and then determines whether this sequence is fake; if the sequence contains more than one fake frame, it will be classified as fake, otherwise it will be classified as normal.

The GAN objective function model has four loss functions, two of which are reconstruction losses, and two are adversarial losses.

1) Reconstruction loss functions. The first one calculates the difference between images, while the second one applies the Laplacian of Gaussian (LoG) on images and then calculates the difference between them.

2) Adversarial loss functions. The first one is an adversarial loss for the frame used to detect whether the frame is fake, and the second one is an adversarial loss for the sequence used to detect whether the sequence is fake.

3) We have used three nonzero weights to balance each of the abovementioned loss functions.

B. Anomaly Detection

The second part of our architecture is a deep learning model that detects anomalies in future frames. This model uses MIL, which begins with dividing the videos into two bags, positive (anomaly) and negative (normal), which render the model a weakly supervised framework that does not entirely need temporal annotation to train but rather needs labels for anomalous videos and normal videos. This solves the problem of preprocessing time where one does not need to indicate the anomalous duration in videos before feeding them into the model. The positive and negative bags contain instances of videos of fixed length; these clips are passed to a feature extraction model (C3D [26], I3D [24], Tube CNN [8]). Each feature extractor has different results; these models are used to extract the required features to be trained by the fully connected deep neural network, which will produce scores. These scores are then used by the MIL's ranking loss function [3] and determine the threshold between the highest scored

anomaly clip and the highest scored normal clip. Thus, our problem changes from a classification to a regression problem because of an insufficient number of anomalies.

IV. Results and Discussion

We used the proposed models with two datasets and performed calculations to analyse how accurate the future frames are.

A. Datsets

The CUHK Avenue Dataset contains 16 training videos and 21 testing ones with a total of 47 abnormal events, including throwing objects, loitering, and running. The size of people may change because of camera position and angle.

Meanwhile, **the ShanghaiTech Dataset** [15] is an extremely challenging anomaly detection dataset. It contains 330 training videos and 107 testing ones with 130 abnormal events. It consists of a total of 13 scenes and various anomaly types. In this research, we used only 2 of the 13 scenes because of resource shortage.

B. Training Details

The training details of each model are explained in this section with the problems we faced in the training process.

1) Retrospective Cycle GAN: We used the ffmpeg [25] library and methods to remove the duplicate frame and reduce the time elapsed by the video. We adjusted three frames per second to summarise the videos and made a list of clips, each consisting of five frames. We normalised the intensities of all input sequences N from -1 to 1 and flipped the input sequence frames horizontally with a probability of 30% for data augmentation. We used the Adam optimiser [10] for the mini-batch stochastic gradient descent method and a batch size of 5. We resized the frames to 256x256 as the default input for the model.

2) Multi-Instance Learning: We split the dataset videos into clips at 16 frames each. Then these clips were fed into a C3D pretrained model on the Sports-1M dataset. The features were then segmented by empirically set segments (32) and passed to a fully connected layer (FC). This network produced scores that will be used by the MIL's ranking loss function [3] and determine the threshold between the highest scored anomaly clip and the highest scored normal clip. Therefore, our problem changed from a classification to a regression problem because of the insufficient number of anomalies.

Training Problems

- The generation model needed a considerable amount of GPU memory, which was a problem at first, as no machines were available to perform such task. However, we tried using Google Colaboratory, as it offers a 12 GB VRAM GPU, which was sufficient for training the model. Finally, the model was trained and produced some initial results, but the GPU was not as fast as expected, so we managed to distribute the work to five individuals to at least get an acceptable number of epochs.
- There were no available datasets as expected, so we researched this problem and found the two datasets that we adopted and used them for several reasons. One is that these datasets are lightweight in size and of high quality, so the normalisation and preprocessing would not destroy the data as was the case with normalisation and the UCF-Crime Dataset.

C. Evaluation Metrics

1) Retrospective Cycle GAN: The predicted frames were evaluated using three metrics that are frequently used for video prediction: mean squared error (MSE), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR). Lower is better for MSE, while higher is better for PSNR and SSIM.

2) Multi-Instance Learning: We used the frame-based receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) to evaluate MIL performance, as the ROC curve shows the trade-off between sensitivity or true positive rate and specificity or complementary false positive rate. Curves closer to the top-left corner show better results.

D. Results

The results were split into three sections for clarification, as our model was split into two separate models and then integrated into a single framework.

1) Future Frame Results: . Considering the current runs that have been done by the current resources from Google Colaboratory, we worked on future frame prediction, which is the first phase of our pipeline, and trained it on different datasets with different FPS methods. Our results are shown in Table 1.

TABLE I
Results on Future Frame Predictions

Dataset	*FPS*	*Epochs*	*SSIM*	*PSNR*	*MSE*
Chunk Avenue	5	21	0.9424	29.40	1.41
Chunk Avenue	10	50	0.9715	35.47	0.44
ShanghaiTech	3	259	0.9528	31.65	1.02

2) Anomaly Detection Results: We tested the model on the Avenue Dataset and achieved an accuracy of 60.47% (Fig. 3); for the ShanghaiTech Dataset, we achieved an accuracy of 88.27% (Fig. 3). Results are shown on the ROC curves.

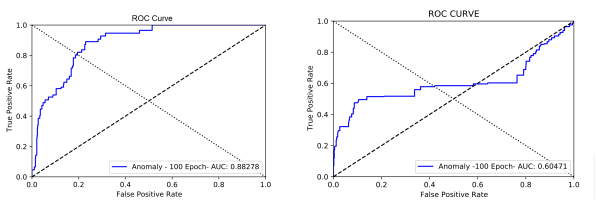


Fig. 3. ROC curves on anomaly detection: The left curve corresponds to the ShanghaiTech Dataset; the right curve corresponds to the Avenue Dataset

3) Proposed Integration Model Results: We used the trained generator from the retrospective cycle GAN; for every four ground-truth frames, three frames were generated and then combined to form a video that contains the generated frames with some ground-truth frames. The clips produced from this were used by the anomaly model to produce an ROC curve (Fig. 4), showing its accuracy. In the Avenue Dataset, the accuracy improved by 8.471% to 68.94% compared with the anomaly detection model with no future frame prediction, while the ShanghaiTech Dataset showed no noticeable improvement at 88.26%. We proved that we can depend on the future frames generated from the GAN model for anomaly prediction.

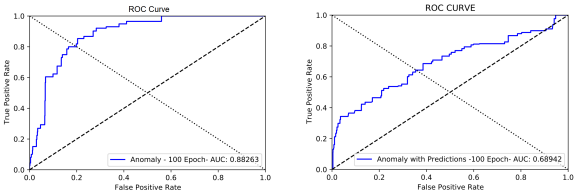


Fig. 4. ROC curves on anomaly prediction: The left curve corresponds to the ShanghaiTech Dataset; the right curve corresponds to the Avenue Dataset

V. Conclusion

In summary, we proposed a framework integrating two models: a future frame prediction method and a weakly supervised deep learning classification method. The framework was used to predict anomalous behaviour in surveillance cameras to prevent crime. We used the future frame prediction (GAN) method to close the gap left by insufficient anomaly prediction methods. By predicting future frames containing anomalous behaviour before it happens by a certain duration, we showed the possibility of depending on the predicted frames, as accuracy either increased or did not change using the given datasets. Our results not only open possibilities in anomaly prevention but also provide a good understanding of what could be done with new improvements in future frame generation methods. In future works, we are looking forward to modifying GANs to produce a higher number of frames with higher SSIM to save as much time as possible to predict future anomalies much earlier and improve the deep learning method to achieve higher accuracy using the given datasets.

References

[1] Mladen Adamovic. Crime in egypt (url: www.numbeo.com).

[2] Mohamed Chaabane, Ameni Trabelsi, Nathaniel Blanchard, and Ross Beveridge. Looking ahead: Anticipating pedestrians crossing with future frames prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), July 2009.

[4] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In Fengyu Cong, Andrew Leung, and Qinglai Wei, editors, *Advances in Neural Networks - ISNN 2017*, pages 189–196, Cham, 2017. Springer International Publishing.

[5] T. Ganokratanaa, S. Aramvith, and N. Sebe. Anomaly event detection using generative adversarial network for surveillance videos. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1395–1399, 2019.

[6] P. Gujjar and R. Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2097–2103, 2019.

[7] J. Honovich. Is public cctv effective? (url: https://ipvm.com/reports/is-public-cctv-effective/).

[8] Rui Hou, Chen Chen, and Mubarak Shah. Tube convolutional neural network (t-cnn) for action detection in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Jasmeen Kaur and Sukhendu Das. Future frame prediction of a video sequence, 2020.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[11] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[12] Y. Lai, R. Liu, and Y. Han. Video anomaly detection via predictive autoencoder with gradient-based attention. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

[13] Federico Landi, Cees G. M. Snoek, and Rita Cucchiara. Anomaly locality in video surveillance, 2019.

[14] Alex X. Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction, 2018.

[15] W. Liu, D. Lian W. Luo, and S. Gao. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning, 2017.

[17] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. 2013.

[18] Y. Lu, K. M. Kumar, S. s. Nabavi, and Y. Wang. Future frame prediction using convolutional vrnn for anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8, 2019.

[19] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[20] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[22] D. Roy, T. Ishizaka, C. K. Mohan, and A. Fukuda. Vehicle trajectory prediction at intersections using interaction based generative adversarial networks. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2318–2323, 2019.

[23] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[25] S. . Tomar. (2006) converting video formats with ffmpeg. linux journal, 2006(146), 10. url: https://ffmpeg.org/about.html.

[26] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.