ANOMALY DETECTION IN VIDEO
GAN BASED APPROACH

A thesis submitted

for the Degree of

MASTER OF TECHNOLOGY
IN
INFORMATION TECHNOLOGY

by

Prabhakar Kumar IRM2017008

UNDER THE SUPERVISION OF
Dr. Krishna Pratap Singh
IIIT-ALLAHABAD

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD
(A UNIVERSITY ESTABLISHED UNDER SEC.3 OF UGC ACT, 1956 VIDE NOTIFICATION NO. F.9-4/99-U.3 DATED 04.08.2000 OF THE GOVT. OF INDIA)

A CENTRE OF EXCELLENCE IN INFORMATION TECHNOLOGY ESTABLISHED BY GOVT. OF INDIA

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

ALLAHABAD

(A Centre of Excellence in Information Technology Established by Govt. of India)

CANDIDATE'S DECLARATION

I, do thusly pronounce that the work displayed in this proposition entitled **Anomaly Detection in Video : GAN based approach**, submitted in the fractional satisfaction of the level of Dual Degree (B.Tech+M.Tech), in information Technology at Indian institute of information Technology, Allahabad, is a bona fide record of my unique work completed under the supervision of **Dr. Krishna Pratap Singh**, due affirmations have been made in the content of the proposition to all other material utilized. This theory work was done in full consistency with the prerequisites and limitations of the endorsed educational modules.

Date: / /

Place : Allahabad **PRABHAKAR KUMAR (IRM2017008)**

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
ALLAHABAD

(A Centre of Excellence in Information Technology Established by Govt. of India)

CERTIFICATE FROM SUPERVISOR

I suggest that the proposition work arranged under my/our supervision by **Prabhakar Kumar** titled **Anomaly Detection in Video : GAN based approach** be acknowledged in the halfway satisfaction of the necessities of the level of Master of Technology in information Technology for Examination.

Date : / /

Place : Allahabad

(Dr. Krishna Pratap Singh)

Counter Signed By Dean (Academics)

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

ALLAHABAD

(**A Centre of Excellence in Information Technology Established by Govt. of India**)

CERTIFICATE OF APPROVAL

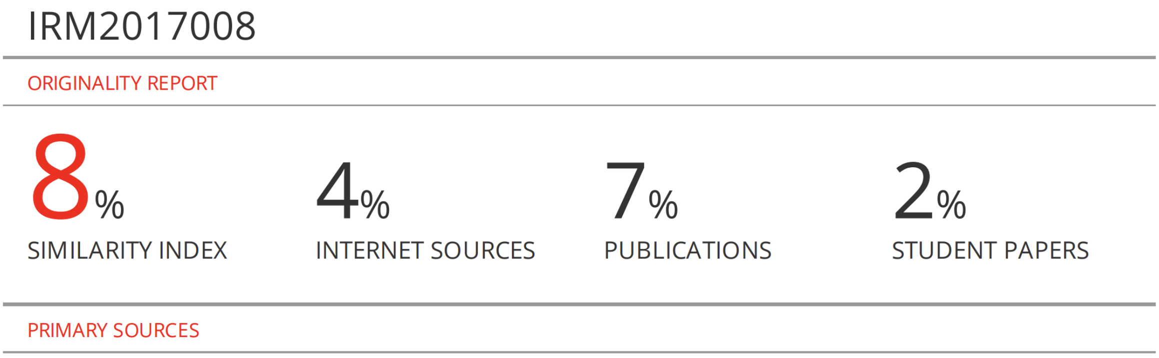
The forgoing thesis is hereby approved as a credible study in the field of Information Technology carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it is submitted.

Signatures of the Examiners Committee Members

(On final examination and approval of the thesis)

TURNITIN ORIGINALITY REPORT

The content in this document was found to be 8% plagarized as reported by the Turnitin plagiarism checker.



IRM2017008

ORIGINALITY REPORT

8%	4%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

Submission Author: Prabhakar Kumar

Submission ID: 1881977859

Submission Date: 13-Aug-2022 12:50AM (UTC-0400)

Word Count: 8667

This is to certify that the thesis titled **”Anomaly Detection in Video : GAN Based Approach”** has been checked for plagiarism

Dr. Krishna Pratap Singh

Prabhakar Kumar (IRM2017008)

Acknowledgements

I would like to express my sincere gratitude to all those who provided me with resources and guidance to complete my thesis.

I am profoundly grateful to **Dr. Krishna Pratap Singh** for his expert guidance and continuous encouragement throughout to see that this thesis meets its target since its commencement to its completion. His instrumental contribution in the proposed work has made it conceivable. I would like to express my deepest appreciation towards all my professors throughout, as their teachings and guidance was clinical in paving the way for a research oriented overlook over problems, with which I am able to complete my work. I sincerely thank the Director of IIIT Allahabad for providing a research oriented infrastructure.

I would like to express my sincere heartfelt gratitude to all the staff members of Information Technology Department who helped me directly or indirectly during this course of work.

Finally, I would like to thank my family, my friends who have had my back every single time and without whose support I would have not been able to reach this point in life.

Prabhakar Kumar

ABSTRACT

Anomaly detection in videos refers to recognition and identification of tasks and events happening in the context of video that does not follow the normal or regular behavior, or show considerable variation from it.[1] Anomaly Detection has been explored by researchers over wide spectrum of applications, using various approaches, but with the efflux of Deep Learning techniques and availability of higher computational resources, there has been a surge in task of anomaly detection in any sort of visual data, primarily video data, owing to its applications such as in video surveillance and medical patient monitoring systems.

In this paper, I propose a framework for detection of anomalous events in video data, based on Generative Adversarial Networks(GANs)[2][3]. My proposed framework is composed of two different networks that work together and are trained adversarially in a Self-supervised manner, as characterized by a Generative Adversarial Network. The generator part of the framework works as an image-to-image translator which tries to predict or impaint the future frame removing the irregularities or any anomalous behavior it identifies, while the discriminator part of the framework tells how well the generator was able to impaint the future the frame by giving out a score based on possibility of an anomalous event. The generator tries to fool the discriminator by predicting as better a frame without anomaly as possible, while the discriminator tries to detect any possible oversight.

The framework was designed keeping in mind to keep the model as computationally inexpensive as possible, without having to compromise much with the work efficiency, because of which it can be trained faster and with considerably less computational resources than many of the pre existing frameworks with similar approaches. The proposed framework has been tested over three datasets, and it has been showing promising results given the lower resources it had been trained upon, though the results are not very close to the baseline models.

Contents

List of Figures

Chapter 1

INTRODUCTION

Anomaly detection in video refers to the identification of events that do not conform to expected behaviour. It finds many applications throughout a wide range of use-cases, surveillance video monitoring and medical patient monitoring being two prominent fields. In surveillance video, based upon traffic monitoring, anomalous events refer to events such as overspeeding vehicles, vehicles moving in wrong directions, any accident and others. This kind of anomaly detection for traffic and vehicle monitoring has been under wide research owing to its application in autonomous vehicles. Further in surveillance video, based on crowd monitoring, anomalous events refer to events such as any event of sudden segregation or panic dispersion of crowd or people, irregular movement pattern of any individual, or just as in case of pedestrian datasets, vehicles like bicycles or cars moving through the pedestrian lane. A fast and accurate identification of such events can necessarily be of great help and can be of great use for the design of any kind of autonomous or intelligent surveillance and monitoring systems.[1]

However one of the major issue that we face with this is that there ain't any hard definition of an anomalous events, and also even though we may have like a huge amount of video data but we still would find it difficult to find ample data of particular kinds of anomalous events, so as to train any Deep Learning model upon. Because of lack of any proper characterization and also due to a large array of events that may be flagged as anomaly, it becomes really challenging to address this objective with a simple classification model. As a subjective definition to hold on to as an assumption we can safely say that whatever that deviates from the normal everyday activities and events in the area should be considered irregularity or anomaly. [4]

A lot of research has been done on this objective, with a wide range of approaches starting with the exploration of low level visual features like the Histogram of Oriented Gradients and Histogram of Optical Flow of different subjects in the video scenes, for representing the regular scenes. These types of approaches relying on trajectory based features achieved good results but on the cost of requirement of highly complex systems.[5] In these kinds of methods the objective is to learn a dictionary based on visual and motion features and reconstruct normal events with minimal reconstruction error given the history. A reconstruction with errors larger than a defined parameter would be considered as a flag indicating anomaly.

Deep Learning based models primarily those based upon CNN and CNN-LSTM sequence have been largely explored, where the objective is quite similar to as mentioned for the above approach, working on features like Optical-flows, Social Forces and others. They too, given a history of frames, work onto generation of future frames with as little reconstruction error as possible, and when the model sees a large reconstruction error either based on the whole frame or a smaller segment or a part of it, it marks that as an anomaly. When such approaches are trained to aggregate the result over a whole frame then it leads to frame level anomaly detection task, where we can only predict if there are any anomalous event happening in the frame without any sense of localization, whereas if we are able to localize these kind of approach to work on finer segments of the frame then we can actually work onto segmentation of the anomalous events in the ongoing frame of the video.[5]

The need of localization of the result to not only detect the occurrence of anomalous events but also to fin segment the video frames and identification of segments and regions in the frame, where there is considerable probability of anomaly, was recognized and with the idea of this researchers including Sabokrou, and others approached the objective with Encoder Decoder and auto-encoders based models. These types of models not only are capable of detection of objects and events but are also capable of localizing them by segmentation of frames as per the occurrences. Recently with the emergence of Generative Adversarial Networks, a lot of researchers have been successful in generating high performance video prediction algorithms, and also GANs being a generative based model also found interests of researchers for application in anomaly detection, with a perspective of overcoming the scarcity of positive class data points or in easier terms data points denoting anomalous events.[2][3][6][7]

With a similar approach Wen Liu and others, proceed with a Generative Adversarial Network based on Auto-Encoder based Generator that they trained with a Fully Convolutional discriminator, trying to predict the future frame with as less reconstruction for normal events as possible, while with higher reconstruction errors at certain localized region of the reconstructed frame, they succeeded to detect the anomalous events.[8]

Using the work of Wen Liu and others, as a baseline I propose certain changes in the model architecture simplifying a lot of proceedings, enabling it to work on frameworks with much less computational resources than the authors, like the Google Colab. The authors had worked with a U-Net based network architecture coupled with FlowNet, for the task of efficient reconstruction of future frames, considering not only the spatial and temporal features but also the optical flow of objects, as facilitated by the FLowNet. With training over video samples denoting normal events, the generator would predict future frames with normal events without much error while for anomalous events there would be regeneration errors, like distortion and blurring of reconstructed frames, which they used to identify anomalous events. [8]

With the proposed changes, my generator model works more like an Impainter trying to predict the future frame impainting the anomalous events, while the normal events are supposed to be transmitted through frames without any or much distortion. The discriminator on the other hand is almost similar to that as the base model, which is trying to predict the measure of how the prediction is done by the generator, with the parameter of likelihood of an anomaly being there in the predicted frame. Once the generator is successfully and sufficiently learnt, given a series of continuous video frames, it would successfully predict the immediate future frame, impainting any occurrence of the anomalous events. The model was also simplified to enable better robustness by the use of pixel based average of frames over history, rather than stacking up of previous frames, as suggested by Mohammad Sabokrou and others, as this would supposedly increase the robustness of the model without incurring much loss to the accuracy of detection. The trained model was then tested first on IR-MNIST dataset, where it was supposed to detect occurrence of digit 3 as anomaly and on UCS Pedestrian Dataset 1 and 2, where it was supposed to detect anomalous events like vehicles and bicycles, passing by a pedestrian path, and given the resource it did perform well though not giving as high accuracy than the baseline model.

Chapter 2

MOTIVATION

Anomaly detection finds a varied set of applications across myriad domains, and have been an integral part of many on-working systems today. A very simple and straightforward application being known to Machine Learning and Deep Learning, would be to find anomalies or better called outliers that we come across in datasets. Such anomaly detection or better called cleaning of dataset, results in a very significant boost in performance of training models.

With the advent of the internet and more so with the digitization of money and payment methods, came the fraudulent transactions and identity thefts. It was noticed with statistical analysis that the fraudulent transaction carried out using identity thefts and data sniffed from multiple resources followed a pattern that varied significantly from the actual ones. This was where anomaly detection found its significant application in fraud detection. The same underlying approach can also be seen to have been applied in network systems across the companies, where they use anomaly detection techniques for intrusion detection in their private systems and networks. Anomaly detection in time series data also found some breakthrough success in fields like stock marketing.

With the enhancement of capabilities of computational systems, the applications of anomaly detection expanded to fields like images and videos lately. The first few successful attempts were made by military systems, where they used anomaly detection techniques to detect enemies and for monitoring and surveillance purposes. Even around us, we can find the application of anomaly detection being used for traffic monitoring, and also for surveillance at public places by law enforcement departments. Any enhancement in such kinds of systems, be it related to a response in higher resolution space, a faster response or even a system requiring comparatively lesser resources would help a lot in pushing forward the boundaries of application of anomaly detection techniques.

Few more significant applications of anomaly detection that motivates us to research more and more into it, including its application in space technologies and medical diagnosis. Anomaly detection in medical diagnosis machinery has been heavily successful in detecting many problems that may go unnoticed by the human eye. In space technologies as well, the maintenance of space stations and the diagnosis of its condition is heavily done through anomaly detection systems only as human effort is hard to be delivered. Having seen so many applications and endless possibilities of expanding to newer realms of applications motivated me to know more about it and try to contribute with my efforts for the development of anomaly detection techniques.

Chapter 3

BACKGROUND

In this section, we will iterate over the background of anomaly detection, some of the ways that have been used to detect anomalies on many sorts of datasets, and also about GANs and how these were found in a breakthrough introduction in anomaly detection.

Anomaly is any thing that is out of the ordinary behavior from a defined set. When we see this from the point of view of machine or deep learning, where everything is defined in terms of data points, we with the strong belief that every set of defined data points follow a certain behavior, maybe linear, quadratic or some complex analogical behaviors that we are now able to detect through deep learning. When we find some data points or some observations that fall out of the normal behavior or pattern, then we tend to call it as an anomaly, based on some threshold of how much tolerance should be held while flagging some observation as anomaly or not. These observations that do not conform to the expected behavior are synonymously called anomalies, often outliers, discordant observations, exceptions or containments based on the different domains of application.

Something that we must keep in mind while learning about anomalies is noise that is present in datasets. All the datasets that we generate are based on some observations that are expected to carry some mistakes or miscalculations, because of which there are expected errors that get noted along the way. Often in some observations these errors may cause the observatory value to deviate a lot from the actual which may lead to that being treated as an anomaly. This is a reason why on any dataset, we first try to filter out possible noise that can be observed or detected through popular methods like over or under sampling, through some performance metrics that are tolerant to noise. Nowadays even some complex techniques like wiener filtering and many a times in vast applications where there is no lack of resources to play with, RNNs have been under good use for removal of noise.

The advent of anomaly detection finds its way back to the 19th century, where the earliest mentions of detecting anomaly or outliers in data was studied by the statistics community in Edinburgh. Over the years, the significance of anomaly detection has been growing because of its applications being discovered in a wide range of domains, to find significant and often critical information. An application that has been much researched upon and has been under extensive use and often extensive scrutiny, is detection of anomalies in credit card transactions and identity thefts. When the digitization of currency and payments hit the world, it came along with major issues of fraudulent transactions that differed from the actual ones with peoples faking out identity based on some actual data that could be sniffed through some leaking resources. With the advent of anomaly detection for finding anomalous transactions, and also along with the various security levels that have been followed over the internet, we are now able to filter out the anomalous and fraudulent transactions.

There have always been challenges that have been faced by researchers while working with anomaly detection in many fields. One of the major challenges that has been prevalent over the years is the challenge of recognizing and defining the normal regions that encompasses every possible normal behavior while minimizing the inclusion of anomalous ones. Also even if we are able to define the normal region then in a lot of applications, the behaviors keep evolving and change what we may recognize as an anomaly today being defined outside the normal regions, may be good enough to be included in the normal domain some time later. Furthermore another major issue is the lack of availability of labeled data for training and validation of the models. Even many of the available datasets are sets of imbalanced ones, where the occurrence of one set of classes overpower the others, and leads to imbalanced learning.

As of anomaly we can define them in many types that include point anomalies which are the simplest type of anomalies to detect, contextual anomalies where some observations are labeled as anomaly or otherwise based on the context that may change overtime and based on some extrinsic variables, or collective anomalies, in which we defined a set of observations as collective anomaly even though if we look with the point anomaly based approach then often there may be observations in them that are not anomalous.

Coming over to the various techniques that have been in use for anomaly detection, we can broadly classify them into three main categories as follows:

- *Supervised Anomaly Detection* : In supervised anomaly detection techniques, it is intrinsically assumed that the labels classifying normal and anomalous data points are available in the dataset. The overall approach of such models is to build a predictive model for normal vs anomaly class, such that it differs the output in range where one end to the range points to the normal class and the other to the anomalous class. This technique has quite a large resemblance to building traditional predictive models but just that the domains and the inference of information is more complex than the conventional. These types of methods also suffer from two major issues namely Imbalanced learning due to imbalanced class distribution and second is actually with the collections of data and information itself where it is often hard to find correct representative labels for the data points.[9]

- *Semi-Supervised Anomaly Detection* : In the section above we came across the problem of imbalanced class distribution, when we study anomaly detection and their datasets. It basically means that the occurrence of normal classes are way higher in numbers than their anomalous counterparts. So one way to have overcome this problem, that for sure would have lead to imbalance learning is to train the models only on normal data points so that the model is accurately able to detect the normal classes and for the occurrences where the output vary from the expected range of normal classes we can safely label it as anomaly.[10] On the other hand we might also have looked into this the other way around, by using only the anomalous occurrences, but these techniques are commonly not used because it is difficult to compile a self sustained and inclusive set of training dataset encompassing all the possible anomalous behaviors.[9]

- *Unsupervised anomaly detection* : The unsupervised anomaly detection methods work on the assumption that the occurrence of normal classes are way higher in number than the anomalous counterparts, and hence we may not require the labels for any distinguishing effect. Even we may be able to adapt many semi-supervised techniques, just by using a sample of unlabeled training data, but we must keep in mind that these unsupervised techniques suffer a lot from high false alarm rate when the occurrences of anomalous classes rise over some threshold as defined by the tolerance of the method.[9]

Now we are well equipped to be able to discuss the use of Generative Adversarial Networks for the use of anomaly detection. Generative Adversarial Network in an overview is basically a machine learning model where two different models, more likely two different neural networks compete with each other to become more accurate in their respective predictions. The two neural networks where one is named as Generator as it generates samples out of some parameters in order to fool the other neural network which is called the Discriminator whose whole point is to identify the output it receives from the generator and classify it better. The training step in GANs is also quite different from the conventional techniques as it is based on a feedback loop where the which as proceeds results in generator able to give output that resemble the actual data more likely than the previous instances, while the discriminator getting better and better at discriminating and flagging the output of generator as fake.[2][3]

Generative Adversarial Networks have been the go-to state of the art technique when it comes to many tasks related to either image generation, regeneration or painting an image based on some information. Generative Adversarial Networks was introduced to the community by Ian Goodfellow, a then research fellow at Stanford, in the year 2016. Back then he introduced this technique of making two different neural networks compete where one will generate data while the other will work on its validation, and with the loss both will get better at their respective tasks. He in his paper had illustrated a very pixelated image of a human that was generated by the generator, which for the sake of information has been training ever since and now has been able to generate images that actually are too difficult for even humans to tell that they aren't of real people.[2][3]

With the introduction of GANs there seemed to be a flood of information and techniques based over the GANs, and many researchers were quick to adapt the approach by varying the set of neural networks and the loss calculation techniques and came up with new variants of GANs. Some of the more famous types of GANs include DCGAN, Conditional GAN, Cycle GAN, followed by many more that found their expertise in different applications and domains.[11]

A Deep Convolutional GAN, short formed as DCGAN, is very similar to GAN but it specifically focuses on using Deep Convolutional networks in place of fully-connected neural networks. Convolutional networks in general find areas of correlation within an image, or in simpler words look for spatial correlations. This makes DCGAN especially suitable for applications related to image or video based processing. Here we also got introduced to a new approach of convolutional application called the Deconvolution Network which is specifically used as the discriminator in the DCGAN, that internally is nothing more than a transposed convolution operation doing the task of upscaling.

The Conditional GAN is an extension of the GAN architecture, where both the generator and discriminator are conditioned on some extra information other than the image input or the actual input. This extra information can be something like class labels or data from other modality, which hence makes this type of GAN something that takes GANs to semi-supervised or often supervised learning realms. The major drawback in GANs until the advent of CGAN, was that it was able to generate output of a single class that it was trained upon, but by adding extra information mostly based on class labels, GANs were now able to generate images or samples of only one particular class haven been trained on multiple though.

Apart from these two, one major variant of GAN that I am using for my application is CycleGAN. The major application of DCGAN and Conditional GAN is based on predicting images or highly likely data samples originating from random noises. CycleGAN introduces the technique of image to image translation, which works on the task to train the generator network that maps images from a source domain to a target domain.[12] One thing to note is that it learns the mapping from one domain of images to another. In our application we have a set of frames taken from some input where both anomalous and non anomalous events are noted and we pass the incoming frame with a set of past frames to give a correlation of movement and time, so that the generator acts as am impainter and impaints the future frame without the anomaly. On the other hand the discriminatory tries to detect whether to an image given to it, it is able to find any anomalous event or not.[9]

Chapter 4

PROBLEM DEFINITION

To present a deep learning architectural model based on Generative Adversarial Network, that is capable of detecting or localizing anomalous events from images or frames derived from videos, preferably surveillance videos. The proposed model should be light enough to be able to be trained in low configuration without the need of high computation capable resources, without much compromising on the accuracy and speed to detection of anomaly.

Chapter 5

LITERATURE REVIEW

Detection of anomaly in video or any sort of visual data has been widely addressed with different methods including one-class classifiers, outlier detection or anomaly removal methods. Most of these methods have either been based upon learning and reconstruction of normal class and reject everything else. There is a lot of work done in Anomaly Detection especially on its application on Image or video surveillance. I hereby put some light on some of the recent and effective previous proposed methods and research commited.

5.1 Detecting Irregularities in Images and in Video

The work of Oren Boiman and Michal Irani were one of the earliest when it comes to the application of anomaly detection in video samples, which paved the way for others by introducing the various critical applications that it can be used for. Published in 2007 thier research was a follow up of their own work, where they mainly worked on detection of anomalous body postures on a dataset of human poses. Their method was based on learning a set of dictionary or database of patches of human poses that the model sees during training, and while a it is being tested on a query they proposed that if if the pose patch can be formed by an ensemble of the learnt pose patches during the training, they would denote that as normal else it would be denoted as an anomaly.[4]

Their methodology was a Semi-supervised learning based approach as they trained their proposed model only on normal data, or specifically normal images of human posed. The major steps that their proposed methodology comprised of included dividing the training images into smaller chunks and creating a dictionary from these chunks, when a new unseen sample is seen by the algorithm, they would divide the image into smaller chunks following the same strategy as followed during training, and try to recompose the test image from the chunks from the dictionary that they learnt in the training step. If they are able to recreate the image uptill a certain tolerance they would flag the test input as normal and vice versa. [4]

5.2 Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos

The methodology is one of among the unique works for anomaly detection in video data which was addressed with extraction and learning of low-level features like Histogram of Oriented Gradients and Histogram of Optical Flow along with some high-level features like trajectory of objects in scene, in which work of Colque and others, achieved good performance over the UCSD dataset. They used these features and region based nearest neighbor algorithms to find whether a region has high or low correspondence with the nearest neighbor as found by their strategy, which they used to predict the anomaly.[5]

Their methodology comprised of Semi-supervised training of their model over only normal data or normal sets of images and then introducing the model to anomalous images as in the testing steps, where with the use of general concepts, such as orientation, velocity, and entropy, they proposed a novel spatiotemporal feature descriptor, called histograms of optical flow orientation and magnitude and entropy, based on optical flow information. When the events they test upon differ significantly from the normal patterns that they learn during testing, they confirm the happening of any anomalous event. They in their paper also had introduced a new dataset of their own called the badminton dataset.[5]

The major steps in which we can divide their methodology can be read as: 1) finding the difference between two consecutive frames, 2) extracting the optical flow information signifying the information regarding the movement and spatial orientation of objects and components of the scene, 3) finally defining the spatiotemporal description based on the data collected, which they called as the Histogram of Optical Flow Orientation. In order to avoid the formulation of optical flow for each pixel of the image, they also proposed a binary mask based approach for optical feature extraction. They divide the image into small cuboidal sections and using the above computed data, try to find correlation of the current cuboid with the neighbouring ones, and if they do not succeed in finding much correlation with any of the neighbouring cuboids, they flag that as an anomaly.

5.3 Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder

Many of the contemporary works are based on the reconstruction hypothesis, in which the proposed model reconstructs the upcoming frames with minimum reconstruction loss, with an idea that reconstruction of anomalous concepts is nearly impossible using the previous observations. Works by Sabokrou, and many others are based on the hypothesis as mentioned above, where they have used high reconstruction error as a mark of irregularity.[7]

Their proposed methodology introduced two novel cubic-patch-based anomaly detector, one based on auto-encoder reconstructing an input video patch while the other based on the power of sparse representation of the video patch. They built on an assumption that the auto encoder when trained on normal instances, will be capable to reconstruct patched with lower error and deviation rather than the anomalous patches. The auto-encoder tries to learn discriminative features based on gradient descent. Then then also make use of sparse auto-encoders as a compression tool which enables key point detection in the video patches. One of the major benchmark that they achieved with their model was the 120 frame per second detection rate on UCSD dataset, with which the established the efficacy of the method to be used for real time anomaly detection.[7]

5.4 Real-Time Anomaly Detection and Localization in Crowded Scenes

Sobakrou and others in their work, proceeded with not only the recognition and identification of anomalies in video frames but also localization of the detection over smaller regions of the frame. They approached the objective with an Auto-encoder based model, which has been successfully used for recognition and localization of objects and very specifically for visual segmentation tasks. This was one of the first work which showed detection of anomaly at finer level than frame itself, which then later led to even pixel level anomaly detection in visual data, which still is an open problem and lot of research has been ongoing to develop end to end model capable of generating output with high robustness without losing out much on accuracy.[6]

Their proposed methodology introduced a descriptor based similarity metric between adjacent patches for detecting sudden changes in spatio-temporal domains from the frames extracted from the video sequence. It was a feature learning based procedure which though was time-consuming in training but resulted in learnt features that were very discriminative of normal patches, hence resulting in high true positive while keeping a low value of false-positive. They in their methodology represented video patches from multiple views, using both local and global feature sets, and then modelled all normal patches with Gaussian distributions.

They followed three conditions based on which they would flag a patch as anomalous or not, which are as follows :

- The anomalous patches would not follow the same pattern of similarity with the adjacent patches as the normal patches would.
- The anomalous patches would most likely not follow the pattern based on temporal changes as a normal patch would.

- The probability or likelihood of a patch being anomalous is much lower than that of it being a normal patch.

The first of the two conditions form the basis of the local feature characterization while the third encapsulates the global nature of the scene. These two representations are modelled separately and later are fitted into a set of Gaussian distribution. Following that a decision boundary is calculated for each of the models, and the result is combined to reach a decision and since at the end both the local and global spatial based decision is available, localization of the anomaly is also easily achieved.[6]

5.5 Future Frame Prediction for Anomaly Detection – A New Baseline

WIth the rapid growth of application of Generative Adversarial Networks for tasks related to images and video application, many researchers succeeded in using GAN based Adversarial Models for the task of anomaly detection. In this regard Wen Liu and others in their work, presented a new baseline model using the autoencoder based U-Net Network which they coupled with FlowNet, to generate construction of future frames such that the generator results in high reconstruction error for anomalous events and lower for normal ones. Their model was not only robust and effective in predicting future frames but also proved robust to the uncertainty in normal events and the sensitivity to abnormal or anomalous events.[8]

Their proposed methodology was built over the future frame prediction technique where he made use of the adversarial training feature based of GAN. Their generator is so trained during training that it should be able to impaint a future frame given a set of past frame, that only has the non anomalous events, and thus eliminating the anomalous occurrences. This surely will lead to a high error between the predicted frame and the actual frame that we will get, and this will act as the basis based on which we can define a threshold according to the dataset, to differentiate between anomalous and non-anomalous events and frames.

Chapter 6

PROPOSED METHODOLOGY

Base Model :

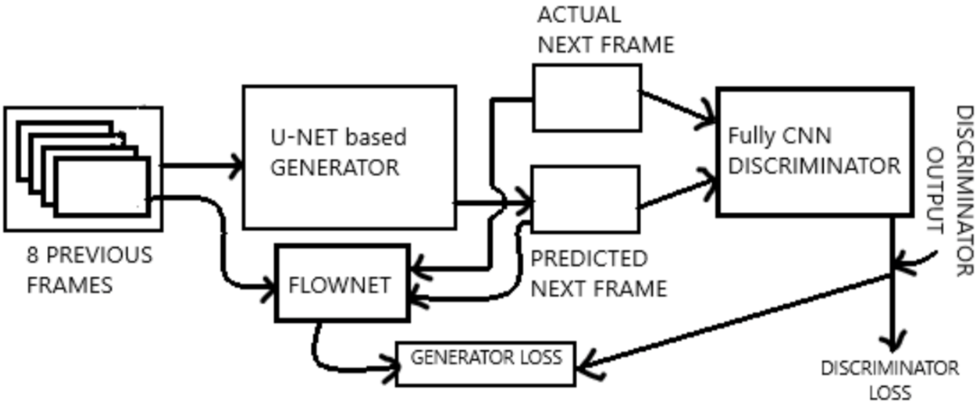


Figure 6.1: A representation of model used by Wen Liu

I have used the work by Wen Liu, as a base model for my own work. The reason for selecting their work as a base and their model, is that they themselves had implemented their model coupling two different Deep Learning models that have been close to state of the art for their objectives. They also had trained their model in adversarial manner as to be done in Generative Adversarial Networks, where in their generator part they have coupled an adaption of U-Net[13] and FlowNet[14] to generate the future predictions of video frames. U-Net is an auto encoder based evolution of Convolutional Neural Networks, developed in 2015, which is majorly based for the task of image segmentation over multiple classes. With a view of detecting the anomaly, a U-Net based model can be purposely modeled to detect and segment anomalies in visual data, either in images or in video frames.[8][13]

The author also used FlowNet[14], which is also a Convolutional Neural Network based architecture capable of estimating the optical flow of objects in image or video frames. As in the input to the generator network they stacked up 8 frames sequentially, say I(t-8) to I(t-1) and used that to predict the next frame, say I(t). This way they can be sure of preserving the intensity and gradient distance between the prediction and actual data, and also to move a step ahead to preserve the temporal coherence they used, the optical flow feature being passed from I(t-1) to I(t), using the FlowNet[8].

The above two components constitute the generator part of the GAN model, which on efficient and sufficient training would be able to predict the future frame keeping all the features in mind. The discriminator they used is a Fully connected CNN based model, which outputs values between 0 and 1, where 0 denoted the fake label while 1 denoted genuine label. If for a frame the cumulative score form the discriminator comes out to be close to 0, then that means that the generator has not been able to predict the future frame as good as it should have, meaning that there must have been an anomalous event, since the generator must be good efficient in predicting the normal events and weakly efficient in predicting the irregular ones.

With the above model setup, the generator and discriminator are adversarially trained in a semi-supervised manner. The objective of the generator is to predict frames so as to fool the discriminator to classify them into class 1. The generator and discriminator are trained in an alternative update manner, like in the step when Generator is being trained the weights of Discriminator remain constant, while on the other hand when the Discriminator is being trained the weights of Generator are kept at constant. The Generator slowly learns to give a more accurate prediction of the future frame, while the Discriminator slowly gets better at predicting the difference between the predicted and the actual frame.

Changes to Base Model :

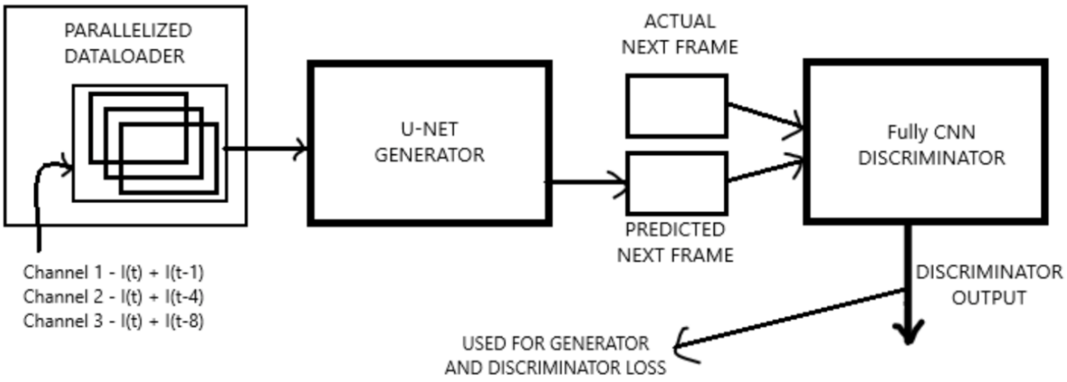


Figure 6.2: A representation of model proposed by me over Wen Liu's

There were a few caveats that I recognised for the application of the base model for my own application and given the resources I had. Firstly the cumulative model required a very high computational power to be trained and large training time for training, but on Google Colab, each user for each session gets a limited amount of resources in terms of RAM and memory and also not a highly-efficient GPU system. While training the base model after adaptation on colab, I was only able to run over 2 successful epochs over the UCSD Pedestrian 2 dataset, after which the output was nothing useful. So, it gave me a suggestion that I need to take steps to make the model faster and less computationally expensive, either by identifying unnecessary complexities or looking for simpler ways to overcome and get going with some features of the dataset, or if required maybe introduce some kind of parallelism in working of the algorithm.

In their work by Sabokrou, as a part of pre-processing step for video annotations, they had suggested that instead for tasks involving previous or historical frame data for visual anomaly detection, instead of stacking up previous frames as was done by Wen Liu, so that later we can either use a LSTM sequence or use 3D convolutional filters, in which one dimension of filter iterates over the data samples of the previous frames, we can simply use pixel-wise average of frames to interpret both the shape and motion features. This way we can significantly reduce the number of channels that the model will be iterating over, and in turn reducing the number of trainable parameters making the model faster and more robust, even with the use of 2D CNN models as in within the U-Net Model. With this strategy I also eliminated the use of FlowNet which further simplified the model architecture. [15]

So as to define the pixel wise averaging scheme, say ye have to predict the I(t+1) frame, given all previous frame with the anomaly impainting task as to be done by generator, we instead of stacking up previous frames, I would take I(t-1), I(t-4) and I(t-8) frames, and use the pixel wise average of all these frames with the I(t) frame, and stack them to input as three channels, much similar to R, G and B channel. This way instead of making the model extensive, I am only making it work equivalent to working with coloured images or coloured video frames. [15]

Furthermore I also equipped the architecture with the DataLoader object as comes with the Python PyTorch library, as it not only keeps the data manageable but also enables multi-process loading of the dataset and simplifies the machine learning pipeline. With these changes in the base model I was able to train the model over the dataset for a much larger epoch iterations than earlier, which demonstrates the increase in robustness of the model architecture.

Chapter 7

DATASET

UCSD Anomaly Detection Dataset :

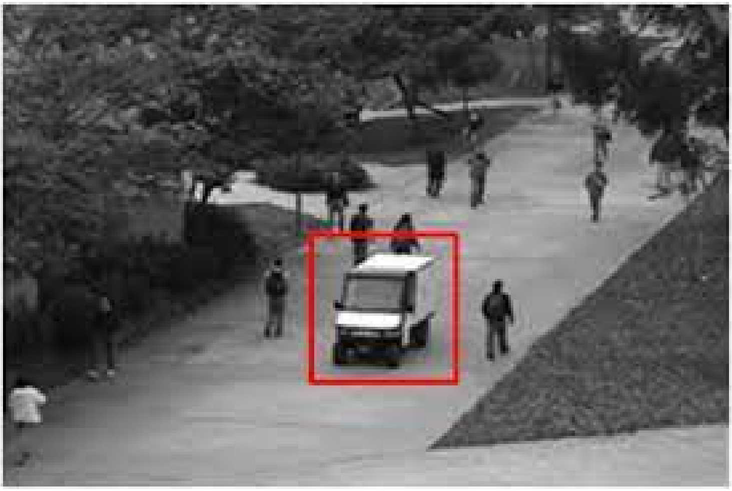


Figure 7.1: A sample scene from UCSD Ped2 Dataset[16]

UCSD Anomaly Detection Dataset[16], which is the most popular of all the available datasets for the task of anomaly detection on visual data. It was acquired with a stationary camera mounted at an elevation, overlooking pedestrian walkways. In normal settings, the video contains only pedestrians, while abnormal events consist of non pedestrian entities like bikes or small carts, passing through the walkways. It also has two subcategories though first in which the camera is oriented parallel to the movement of the crowd and second where the camera is oriented to be vertical to the flow. The total size of the dataset is roughly 700MB, and a lot of variations of it is available for other tasks like segmentation also. (Available at : `http://www.svcl.ucsd.edu/projects/anomaly/dataset.html`)

IR-MNIST Dataset :

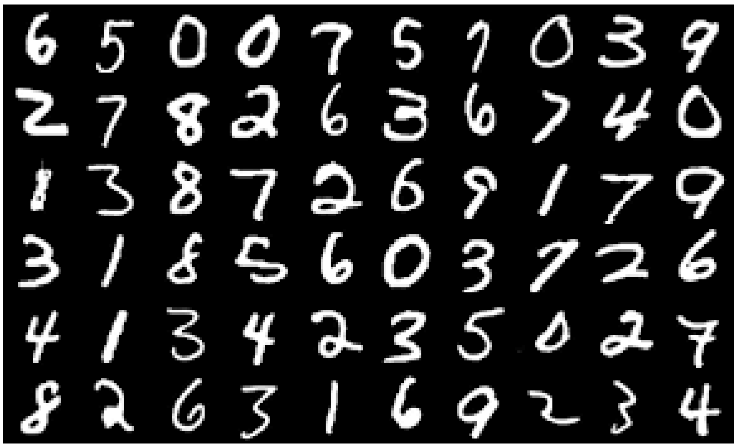


Figure 7.2: A sample from IR-MNIST Dataset[15]

IR-MNIST Dataset which is a toy dataset that was created and used by Sabokrou [15], in which they have created each sample image by randomly selecting 121 samplets from MNIST dataset, and putting them as a 11 X 11 subplot image format. Training data were created without the digit '3', hence training the model over this dataset, it should be able to recognize any occurrence of 3 in the test data, as anomaly and the generator should try to impaint any occurrence of 3 with some other digit. (Avialable at : `http://ai.stanford.edu/~eadeli/publications/data/IR-MNIST.zip`)

Chapter 8

SIMULATION AND RESULT

Deep Learning algorithms are usually ran on computers with high computational powers, preferably having large GPUs, as training over GPUs is very much faster and efficient as compared to normal computations. Generative Adversarial Networks are such deep learning architectures that take a lot of time in training for a satisfactory result. One of the earliest Generative Adversarial Network that draws human face of people who do not actually exist, have ever been training till data and getting better at its objective. In their works, Wen Liu had simulated their GAN architecture over a computer clocked at over 3.40GHz assisted by NVIDIA GE-Force Titan GPU, over which they were able to clock an average frame rate of 25fps.[8]

In my work, I had trained the model over highest available resource over Google Colab with 24GB of RAM and 16GB of GPU memory, and training for 5 epochs on the UCSD Ped1 dataset, took over 3 hours but was not getting satisfactory results. In the later parts, the model architecture was trained over a system with 32GB of RAM and 32GB of GPU storage, and training for 10 epoch took close to 4 hours giving satisfactory results. The current algorithm has an average frame rate of between 11fps-12fps.

For the evaluation of the efficacy of the algorithm and architecture, Receiver Operation Curve (ROC Curve) and the calculated value of AUC (Area under the Curve) by varying the threshold value for regular scores, have been extensively used by researchers over this application. We have also calculated the same, as have compared our AUC score with those of the previous works.

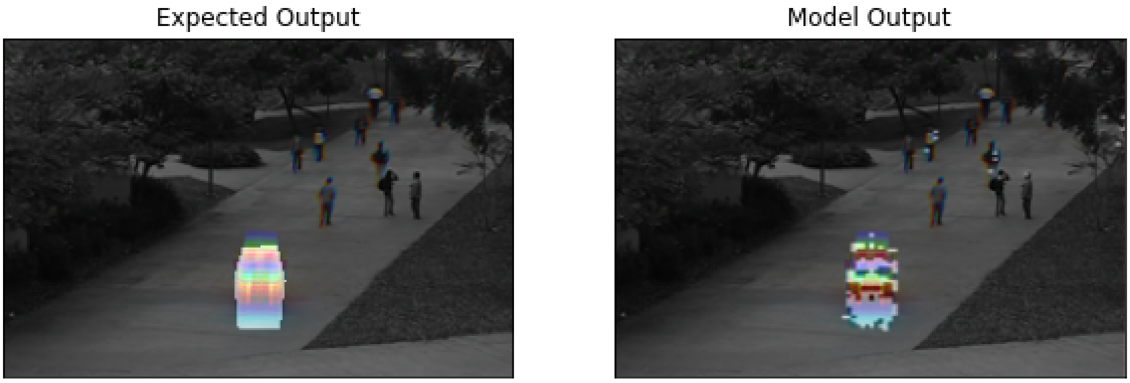


Figure 8.1: Expected Output vs the Model Output : UCSD Ped1

The figure above is a representation of the comparison of the expected outcome and the result given out by the trained architecture. The image on the left is frame with the white masked out region showing the one with anomaly, here a vehicle on the pedestrian track. The image on the right is the outcome of our trained GAN architecture where we can see that the model was closely able to flag the vehicle as an anomaly. Also since in the test scenes we were available with the masks that signified the anomalous data points at the pixel level on the frame, we were not only able to visually judge the models work but also were able to draw accuracy inferences of anomalous points correctly classified, ignoring the non-anomalous points and they were insignificant to out knowledge. The following graph represents the accuracy of the model over identification and prediction of anomalous (excluding non-anomalous) points over multiple frames of one test scene.

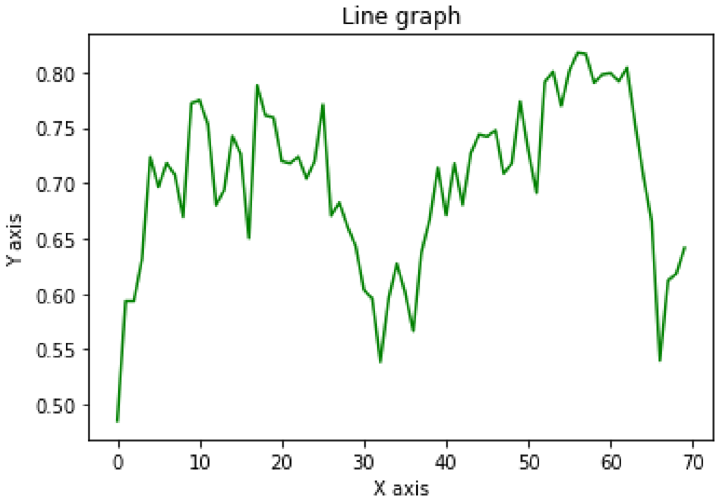


Figure 8.2: Model Accuracy at predicting anomalous points over UCSD Ped1 Dataset

Similar results were obtained over the UCSD Ped2 dataset, but that being a smaller dataset, with the view of not training the model too long to overfit to the training dataset, the model suffers from a larger false negative prediction than it does in the UCSD Ped1 dataset. A possible reason for the aforeshead can also be that this part of the dataset had varying density of crowd over the frames, where in some frames the roads were scarcely populated while in some was densely populated. The following image shows a sample of the comparison between an expected result and the model outcome over a scene from UCSD Ped2 dataset.

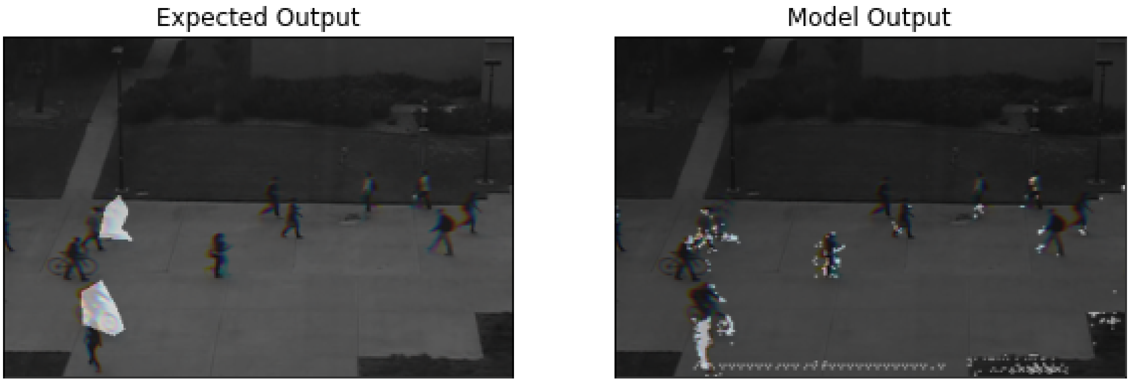
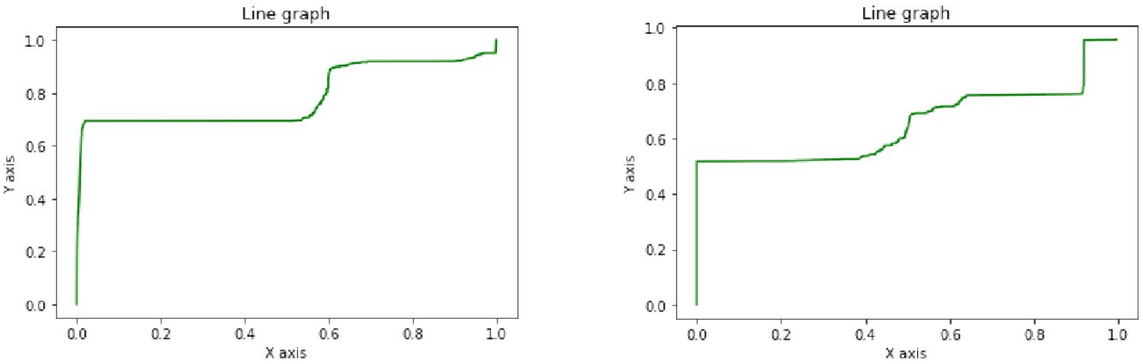


Figure 8.3: Expected Output vs the Model Output : UCSD Ped2

For the overall testing and accuracy score of the proposed methodology, the model was separately trained over each of the UCSD Ped1 and UCSD Ped2 datasets, and were tested over all those test scenes for which we had the anomaly masks available. This facilitated us to evaluate the confusion matrix at each of the frame, based on the pixel level anomaly prediction, and by varying the thresholds for the output of the generator and the discriminator we were able to draw the ROC curve for both the datasets, using which we calculated the AUC score for both the datasets that we comapare with the baseline.

The figure on the left is the ROC curve on testing over the UCSD Ped1 dataset while that on the right is the one on testing over the UCSD Ped2 dataset.

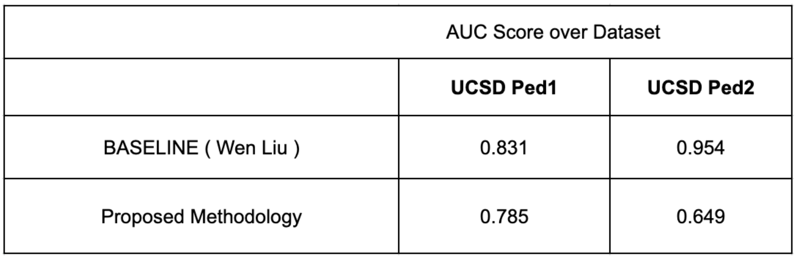


UCSD Ped1

UCSD Ped2

Figure 8.4: ROC Curve over two test datasets

For the final comaprison the AUC score was calculated over both the test scenes over several scenes, and we obtained an AUC score of 0.785 on the UCSD Ped1 dataset in comparison to the 0.831 score of the baseline, but the performance on the UCSD Ped2 Dataset was significantly worse than expected, with an AUC score of 0.649 over the 0.954 of the baseline paper.[8]



	AUC Score over Dataset	
	UCSD Ped1	**UCSD Ped2**
BASELINE (Wen Liu)	0.831	0.954
Proposed Methodology	0.785	0.649

Figure 8.5: AUC Score Comparison

The presence of false positive predictions similar to salt and pepper noise in our model output over the UCSD Ped2 dataset had heavy negative impact over the performance and accuracy. Visually we can see over the comparison of expected output and model output that the model is falsely flagging many non-anomalous points like normal walking person as anomalous when there are several too close to each other, or when they are close to the actually anomalous objects or events like vehicles. The UCSD Ped1 dataset had a consistent density of people over the scenes, and our proposed model is able to perform satisfactorily well in its testing over this dataset, where the outcome is also visually very close to the expected outcome using the anomaly mask.

Chapter 9

CONCLUSION AND FUTURE WORK

In this thesis, a methodology of detecting anomalies in videos has been proposed by predicting the future frame using the data from previous frames. The initial proposal of using time based averaging of frames instead of stacking of frames into layers, to incorporate the information of spatial movements of objects in the frame, has been able to eliminate the need of having a Flownet network to do the same, thus also reducing the overall computational requirement of the model. Further the Generator and Discriminators were also simplified and trained with standard architecture for standardization purposes, with which we were able to train and deploy our model on Google Colab as well. We achieved an AUC score that is close to the baseline work in the UCSD Ped1 dataset, but had a weaker performance in the UCSD Ped2 dataset, but the prediction and accuracy calculations has been done on performance over pixelated data and not only at the frame level as many of the contemporary work has been doing.

A part of future work can be to train and test the model on larger dataset and even on coloured dataset, since the performance drop was seen on the smaller dataset, hence it can be believed that a large dataset would have enough information to better train the model. Furthermore training over coloured or video or even hyperspectral image frames will be a good challenge as well as it would require still higher computation and better thresholding to generate good results.

References

[1] V. Chandola et al. *Anomaly detection: A survey.* URL: https://dl.acm.org/doi/10.1145/1541880.1541882.

[2] Ian Goodfellow et al. *Generative Adversarial Nets.* URL: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[3] Saifuddin Hitawala et al. *Comparative Study on Generative Adversarial Networks.* URL: https://arxiv.org/abs/1801.04271.

[4] Boiman et al. *Detecting irregularities in images and in video.* URL: https://ieeexplore.ieee.org/document/1541291.

[5] Rensso Victor Hugo Mora Colque et al. *Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos.* URL: https://ieeexplore.ieee.org/document/7778165.

[6] Mohammad Sabokrou et al. *Real-time anomaly detection and localization in crowded scenes.* URL: https://ieeexplore.ieee.org/document/7301284.

[7] Mohammad Sabokrou et al. *Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder.* URL: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/el.2016.0440.

[8] Wen Liu et al. *Future Frame Prediction for Anomaly Detection-A New Baseline.* URL: https://ieeexplore.ieee.org/document/8578782.

[9] Abdul Jabbar et al. *A Survey on Generative Adversarial Networks: Variants, Applications, and Training.* URL: https://dl.acm.org/doi/abs/10.1145/3463475.

[10] Lukas Ruff et al. *Deep Semi-Supervised Anomaly Detection.* URL: https://arxiv.org/abs/1906.02694.

[11] Alec Radford et al. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.* URL: https://arxiv.org/abs/1511.06434.

[12] Yongyi Lu et al. *Attribute-Guided Face Generation Using Conditional CycleGAN.* URL: https://openaccess.thecvf.com/content_ECCV_2018/html/Yongyi_Lu_Attribute-Guided_Face_Generation_ECCV_2018_paper.html.

[13] Huimin Huang et al. *UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation.* URL: https://ieeexplore.ieee.org/abstract/document/9053405.

[14] Eddy Ilg et al. *FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks*. URL: https://openaccess.thecvf.com/content_cvpr_2017/html/Ilg_FlowNet_2.0_Evolution_CVPR_2017_paper.html.

[15] Mohammad Sabokrou et al. *AVID : Adversarial Visual Irregularity Detection*. URL: https://arxiv.org/abs/1805.09521.

[16] Antoni B. Chan et al. *Modeling, clustering, and segmenting video with mixtures of dynamic textures*. URL: https://www.researchgate.net/publication/5483792_Modeling_Clustering_and_Segmenting_Video_with_Mixtures_of_Dynamic_Textures.