Dynamic Image Crowd Representations for Improved Anomaly Detection using Generative Adversarial Networks

Samar Mahmoud

A thesis submitted in partial fulfilment of the requirements of the University of Greenwich for the Degree of Doctor of Philosophy

May 2021

DECLARATION

I certify that the work contained in this thesis, or any part of it, has not been accepted in substance for any previous degree awarded to me or any other person, and is not concurrently being submitted for any other degree other than that of Doctor of Philosophy which has been studied at the University of Greenwich, London, UK.

I also declare that the work contained in this thesis is the result of my own investigations, except where otherwise identified and acknowledged by references. I further declare that no aspects of the contents of this thesis are the outcome of any form of research misconduct.

I declare any personal, sensitive or confidential information/data has been removed or participants have been anonymised. I further declare that where any questionnaires, survey answers or other qualitative responses of participants are recorded/included in the appendices, all personal information has been removed or anonymised. Where University forms (such as those from the Research Ethics Committee) have been included in appendices, all handwritten/scanned signatures have been removed.

Student Name: Samar Mahmoud

Student Signature:

Date: 25 May 2021

First Supervisor's Name: Yasmine Arafa

First Supervisor's Signature:

Date: 25 May 2021

Second Supervisor's Name: Cornelia Boldyreff

Second Supervisor's Signature:

Date: 25 May 2021

ACKNOWLEDGEMENTS

Throughout the writing of this thesis I have received a great deal of support and assistance.

First and foremost I am extremely grateful to my supervisors, Dr. Yasmine Arafa, Prof. Cornelia Boldyreff, and Prof. Jixin Ma for their invaluable advice, continuous support, and patience during my PhD study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. It is their kind help and support that have made my study and life in the UK a wonderful time. I would also like to express my appreciation of Professor David Marshall, Cardiff University and Professor Bernie Tiddeman, Aberystwyth University for their time and guidance in my final PhD examination.

In addition, I would like to thank my sister for her wise counsel and sympathetic ear. You are always there for me. I could not have completed this thesis without your support, you provided stimulating discussions as well as happy distractions to rest my mind outside of my research.

To all my friends, thank you for your understanding and encouragement in my many, many moments of crisis. Your friendship makes my life a wonderful experience. I cannot list all the names here, but you are always on my mind. Without your tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my study.

Finally, I would like to express my gratitude to my parents, for their valuable guidance throughout my studies. You provided me with the tools that I needed to choose the right direction and successfully complete my thesis.

ABSTRACT

Crowd formations are inevitable in many environments, and hence planning for, and managing crowds are integral parts of city and event planning. Effective analysis of crowd behaviour and anomaly detection has the potential for more efficient management and is a building block for smart environments. Closed-Circuit Televisions (CCTVs) capture vast footage and are an important information source, some of which contain images of crowds of high density. However, relying on the typical manual surveillance systems for detecting anomalies (any behaviour outlying from established normalcy) in crowds presents complications concerning accuracy and computation power. This research intends to advance the automation of anomaly detection within medium and high-density crowds. Using crowd behaviour analysis methods, anomaly detection is applied to recognise occurrences of anomalous behaviour within crowds. An anomaly within the behaviour of the crowd is detected by analysing crowd footage with the use of deep vision algorithms. Results obtained from the processing of video data can be used to understand the overall scene and discriminate between normal and abnormal behaviour within a crowd.

Application of crowd anomaly detection has improved recently, however, the algorithms currently being used are usually time-consuming, computationally heavy, or require high power consumption. Amongst the work reviewed, both handcrafted approaches, as well as a variety of neural network approaches suffer from a lack of a definition of what "abnormal" behaviour is. Benchmark datasets used to train/test these methods lack sufficiently rich enough data to define anomalous behaviour. Therefore, abnormal events are considered as any events that deviate from the defined normal. Furthermore, state-of-the-art methods also present limitations of applicability to high-density crowds. High-density crowds are not targeted as much due to their difficulty in application. A key contribution of this research addresses this issue with the creation of a public anomalous high-density crowd dataset. The high-density dataset named Abnormal High-Density Crowd (AHDCrowd) has been utilised in training and testing the state-of-the-art crowd anomaly detection methods to evaluate their anomaly detection performance on high-density crowds.

Another key contribution of this research is a novel approach to crowd behaviour anomaly detection. Various dynamic image representations are used as an alternative to optical flow extractions for temporal development features extraction. The features are used in conjunction with image-to-image translation using CGANs (Conditional generative adversarial nets) for anomaly detection within crowds, and the proposed framework is evaluated on benchmark datasets as well as the AHDCrowd dataset. The applied experiments evaluate the effectiveness of utilising various types of dynamic image representation for crowd anomaly detection. The experimental results obtained have demonstrated the efficacy of this approach compared to the state-of-the-art crowd anomaly detection methods.

CONTENTS

TABLES

FIGURES

1 Introduction

This chapter aims to provide a higher understanding of this research topic and detail the motivation, focus, and contributions of this research. The first section details the main motivations behind this research. The next section is a detailed overview of crowd behaviour analysis within computer vision, including the various subfields it contains. The third section includes real-world applications of crowd analysis. Following this, the core research focus and proposed contributions are detailed as well as the methodology of this research. Lastly, the publications of this research and outline of this thesis are detailed.

1.1 Motivation

Surveillance systems have been utilised to ensure public safety, fight and prevent crimes, and prevent antisocial behaviour and nuisances. These systems help monitor crowded venues such as malls, airport terminals, sporting arenas, and concert halls. Surveillance of behavioural changes within crowds in these venues can prevent undesired or even dangerous incidents from occurring. It can also help with the planning and management of crowds in the aforementioned venues. Chaotic activities are usually triggered by abnormal events such as fires, dangerously loud noises, gas escapes, etc. The resulting chaotic behaviour can lead to actions that are just as threatening as the incident itself (Grant and Flynn, 2017). To manually identify/interpret irregular or dangerous incidents is practically impossible (Cao et al. 2009; Joshi et al. 2019). This is because the number of surveillance cameras tremendously exceeds the number of personnel and viewing monitors. Since potential mistakes such as personnel overlooking important incidents may arise from this, surveillance systems must detect noteworthy events on and off screens in an automated manner.

With the use of computer vision systems, crowded scenes can be analysed and studied to interpret a crowd's behaviour and aid in the management of crowded venues. However, analysing crowd behaviour presents difficulties that have prompted further research within the field. Such problems related to recognition, tracking, and motion estimation of crowded scenes. Computer vision techniques brings additional problems such as occlusion handling, self-occlusions, irregular motion direction, and ambiguities (Dee and Caplier, 2010; Li et al., 2015). Furthermore, a crowd of people is often goal-focused and demonstrates both dynamic and psychological characteristics and finding a fitting level of granularity to model the changing aspects of a crowd is complex. Additionally, to construe what is considered abnormal behaviour in a crowd is a computer vision problem. However, with the use of deep vision algorithms, results obtained from the processing of video data can be used to understand the overall scene as well as discriminate between normal and abnormal behaviour (any behaviour outlying from established normalcy) within a crowd.

1.2 Overview: crowd behaviour analysis in computer vision

Automatic analysis of crowd (a large number of people that have gathered in the same location) behaviour is increasingly becoming an important domain in computer vision due to its wide implications on crowd safety and security. Crowds can have different densities (number of people per square meter) such as low, medium and high-density. In images, low-density crowds shows coarse textures, whereas high-density crowds show fine textures. Crowd formations are present in streets, public events, concerts, airports, religious pilgrimages, marathons etc. These venues are vulnerable to many harmful incidents including crowd disasters. Video surveillance has been increasing in many environments to enhance security and prevent disastrous situations. Consequently, vast amounts of data are generated from multiple sources and are increasingly overwhelming surveillance operators. The automation of crowd behaviour understanding requiring limited human supervision/intervention is essential to enable smarter and safer environments. To achieve this, data is extracted from surveillance footage using computer vision methods and technologies to understand a crowd's behaviour automatically.

Computer vision methods are applied to many fields such as autonomous vehicles, healthcare and facial recognition, among others. The fundamental aim of computer vision is to extract high-level information from images and videos. Computer vision tasks such as object detection, classification and localisation, and instance and semantic segmentation are required to extract this information. The general focus of this research is the evolution of computer vision methods that can be applied in crowd analysis. To analyse a crowd, global scene features are extracted from images or videos. Examples of these features include, among others, the number of people in a crowd, trajectories of a crowd and behaviour classification. Computer vision methods in crowd analysis and crowd behaviour analysis are generally categorised into:

- Crowd Counting: An approximation, extracted from an image, of the true count of people in a crowded environment. The approximation is represented as an integer value (Rodriguez et al., 2011a; Gao et al., 2020).
- Crowd Density Estimation: Similar to crowd counting, crowd density estimation is an estimation of the crowding level in an image represented by a discrete value (0-N) (Rodriguez et al., 2011a; Gao et al., 2020).
- Crowd Tracking: The process of tracking an object or person in a crowd throughout multiple video frame sequences (Salim et al., 2019; Shehzed et al., 2019).
- Person Re-identification: Recognising the same object or person across multiple disjoint

cameras throughout different times (Mazzon et al., 2012; Ye et al., 2020).

- Crowd Behaviour Recognition: Analysing a crowd to recognise and classify the collective behaviour of the crowd (Bertini et al., 2012; Matkovic et al., 2019).
- Crowd Behaviour Anomaly Detection: Detecting the collective behaviour of a crowd to determine the level of abnormality presented. Abnormal behaviour of a crowd is defined as any behaviour outlying from established normalcy (Popoola and Wang, 2012; Tripathi et al., 2018).

This research investigates the aforementioned computer vision tasks applied in computer vision for the analysis of crowd behaviour. However, the main aims and contributions of this research are focused on crowd behaviour anomaly detection.

1.3 Applications of Crowd Analysis

Behaviour analysis of crowds can be beneficial but challenging in many fields of application (Li et al., 2015; Zhan et al., 2008). The impact of the research in this thesis would mainly benefit the surveillance and crowd management disciplines, but the methods and algorithms to be discussed are implementable in other applications. Other crowd analysis applications include:

- *Crowd Management:* Public safety is always a challenge in any mass gatherings and to avoid potential catastrophic events, such as overcrowding or bottlenecks, crowd behaviour analysis can be used to determine and apply the best crowd management strategies (Zhan et al., 2008; Lamba and Nain, 2017; Joshi et al., 2019).
- *Public Space Design:* Public spaces such as train stations, buildings, and universities/schools (Li et al., 2015) require specific guidelines on how to be built safely while maintaining the building space efficiently. Crowd analysis can help plan the structural layout for maximal optimisation (Lamba and Nain, 2017; Joshi et al., 2019).
- *Virtual Environments:* Organising and planning events can be enhanced by the use of virtual crowd phenomena in an environment. Crowd analysis can also improve virtual modelling of dangerous conditions and predict how the crowd would react (Grant and Flynn, 2017). This can help prevent the occurrences of potentially dangerous situations (Joshi et al., 2019).
- *Security and surveillance:* Video surveillance is used in many public spaces, some of which with highly crowded scenes. In these situations, relying on typical surveillance systems presents complications concerning accuracy and computation (Li et al., 2015). For real-time detection of a specific event within a crowd, surveillance operators will be

constantly required to observe the scene. Additionally, the number of operators will have to increase to keep up with the number of situated cameras. As for the detection of past events, the footage will have to be stored with acceptable quality (quality that is suitable for the extraction of chosen features based on the models requirements), requiring footage compression and decompression. This is an avoidable computational increase. Operators will still have to go through the footage to detect targeted events manually. With the use of an automated system combined with crowd analysis, extraction of specific actions can be used to alert if an anomaly or irregular action has transpired (Sjarif et al., 2012; Lamba and Nain, 2017; Joshi et al., 2019).

- *Intelligent Environment:* Crowd analysis is a great benefit to creating an adaptive intelligent environment. When a large crowd is gathered in a venue similar to a museum or an art gallery, smart decisions are made to determine where to direct a crowd or if they should be dispersed. These smart decisions can be assisted using crowd analysis based on how the crowd behaves (Junior et al., 2010; Joshi et al., 2019).
- *Entertainment:* The entertainment industry can benefit from crowd analysis by using crowd simulation in divisions such as television, movies, and games. To advance these fields, realistic simulations can be created by understanding how a crowd behaves (Li et al., 2015; Lamba and Nain, 2017).

1.4 Focus of Research

The aim of this research is to improve the performance of current crowd anomaly detection models used in crowd analysis. The advancement of Generative Adversarial Networks GANs has demonstrated its ability to model complex distributions of real-world data. The accurate detection of anomalous behaviour is a challenging task, and a network with a capability of complex modelling can assist in the advancement of this task. Currently, the applications of Conditional GANs (a variant of GANs) for crowd behaviour analysis, particularly anomaly detection, has not been thoroughly investigated. Additionally, dynamic image representations have outperformed optical flow extraction, in the field of action recognition. Optical flow is the typically used motion representation in crowd anomaly detection methods using CGANs and dynamic image representation are an amalgamation of multiple sequential optical flow frames. Therefore, a novel method combining the use of Dynamic Images as motion representations and image-to-image translation using CGANs (Conditional Generative Adversarial Networks) is proposed.

This research also aims to evaluate the performance of state-of-the-art crowd anomaly detection methods in a high-density environment in comparison to low and medium-density crowds. State-of-the-art crowd anomaly detection methods are consistently evaluated on benchmark datasets that only include low and medium-density crowds. High-density crowds are not examined due to the lack of anomalous high-density crowd datasets. To further clarify the focus of this research, the research questions, hypotheses and contributions are described below:

1.4.1 Research Questions

- How can Generative Adversarial Networks for image processing enhance crowd behaviour anomaly detection within medium to high-density crowds?
- What are the associated benefits and trade-offs of utilising the proposed Dynamic Image and CGANs (Conditional Generative Adversarial Networks) method in comparison to the existing state-of-the-art techniques?

To address the aforementioned research questions the hypotheses of this research will be examined in Chapters 4, 5, and 6.

1.4.2 Hypotheses

- As CGANs integrated with optical flow extraction can detect anomalies within medium-density crowds, their application to high-density crowds is expected to be effective.
- The use of dynamic images as an alternative to optical flow will better train CGANs to detect anomalies within medium to high-density crowds concerning accuracy and performance.

1.5 Contributions

This thesis presents a novel approach for crowd anomaly detection by applying Dynamic Images as motion representations and image-to-image translation using CGANs. Additionally, a novel anomalous high-density crowd dataset is created for crowd anomaly detection with highly dense crowds. The main scientific contributions of this research are threefold:

- Generative modelling for anomaly detection in high-density crowds. Conditional Generative Adversarial Networks (CGANs) produces data to a discriminative function to distinguish between normal and abnormal behaviour within medium to high-density crowds.
- The development of a CGAN architecture combined with Dynamic Images (Bilen et al., 2016) provides a novel approach for crowd behaviour anomaly detection.
- A labelled high-density crowd dataset containing normal and abnormal (footage with

anomalous behaviour) has been created. The dataset has been applied to anomaly detection algorithms and has been made public to other researchers.

1.6 Research Methodology

This research's methodological approach is an applied one; it requires both qualitative and quantitative methods. An extensive survey on crowd analysis, behaviour analysis, and anomaly detection has been undertaken to answer the research questions previously mentioned. Quantitative methods were used for data collection. The data is collected from peer-reviewed publications regarding crowds, anomaly detection, crowd behaviour analysis, and generative adversarial networks. Subcategories of each topic are also examined. Data collection is based on novelty, publication-quality, and correlation to the aim of this research. The initial investigation began with crowd analysis, leading to categories such as density estimation and crowd counting, tracking and person re-identification, crowd motion detection and crowd behaviour analysis (Figure 1). Prominent algorithms in each category are explored to further the understanding of each field as well as their influences on each other. The main methods applied as state-of-art were machine learning methods instead of hand-crafted methods.

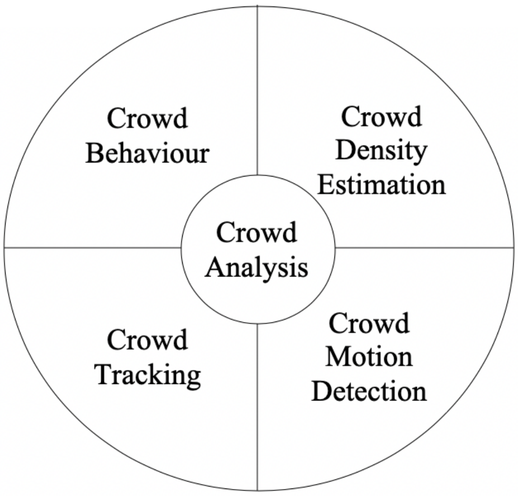


Figure 1: Crowd Analysis categories. Adapted from (Sjarif et al., 2012)

Experimentation with some of the algorithms learnt is documented in Chapter 6. Crowd behaviour analysis, more specifically, detection of anomalies is a well-researched area. However, there are gaps in the application of these methods. Real-world application has not yet been reached due to low performance. Additionally, high-density crowds have not been targeted. This is mainly due to the unavailability of datasets to train and test methods. Continuing this research has led to the novel use of Generative Adversarial Networks (GANs) for anomaly detection within crowds. As an alternative to typical machine learning methods, described in Section 3.3.3. GANs have been explored to utilise their novelty as well as extend the existing work concerning GANs and anomaly detection. The previous work, applied by previous researchers, has demonstrated the ability of GANs to detect anomalies within medium-density crowds with higher performance compared to other state-of-the-art methods.

Appropriate datasets must be determined to evaluate methods and algorithms of crowd behaviour analysis and anomaly detection. Multiple benchmark datasets were found through research, detailed in Section 3.6, but most did not combine features such as high-density crowds, annotations, and the occurrences of anomalous behaviour. To solve this problem, simulation has been taken into consideration. However, crowd simulation was found to be inadequate due to the software's inability to mimic a high-density crowd's behaviour. The complexity of human behaviour, particularly crowd behaviour, surpasses that of simulated behaviour. This has led to footage collection and application of data labelling software shown in Chapter 5.

Experimentation is applied in several fields for this research's objectives, details of the setup, datasets used, algorithms, and results are noted in Chapter 6. Evaluation of the experimental results is noted using both quantitative and/or qualitative evaluation metrics. Some methods utilise various quantitative measures, whereas others lack a solid comparative evaluation metric. This complication has led to the utilisation of qualitative measures to compare the strengths and weaknesses among different methods. The combination of quantitative and qualitative methods for evaluation is the most suitable approach to this applied research.

1.6.1 SEMMA

This research applies the SEMMA methodology to collect data related to crowd behaviour analysis. SEMMA is a data mining methodology consisting of multiple processes, which can also be applied to different aspects of data gathering in different disciplines. The SEMMA process is introduced by the SAS Institute (Goodnight, 2018) and is divided into the following tasks: Sample, Explore, Modify, Model, and Assess. These tasks are usually applied to guide data mining methods. In this research, the methodology supports the data collection process concerning crowd behaviour analysis. The tasks are applied in the following manner:

1. Sample: collecting sample data relevant to crowd behaviour analysis. Collective behaviour, crowd counting, density estimation, tracking, person re-identification, motion representation and anomaly detection are considered in the data collecting process. The research collected regarding each of these aspects is documented in Chapter 2 and Chapter 3.

2. Explore: exploring the sample data is required to gain the required knowledge to successfully complete the contributions this research. Crowd counting and density estimation are an elementary form of crowd analysis, and they define the size of a crowd.

Tracking and re-identification are also forms of crowd analysis that track individuals or small groups from single or multiple fields of view(s).Motion representation of a crowd observes aspects such as the identification of crowd flow, trajectory analysis and dominant motion detection. Lastly, anomaly detection is explored to determine anomaly occurrences within a crowd such as stampedes, persons falling over, unexpected obstacles and surges in the flow change.

3. Modify: the modifying process is applied to narrow down the explored data to find the aim and successively the contribution of this research. The psychological aspects, benchmark methods, handcrafted methods and novel methods are all considered while modifying the data. While many crowd behaviour analysis methods had been subjected to long-term investigations, other aspects are novel and still require further advances. Research gaps are also considered within the modification process; crowd behaviour analysis applied to high-density crowds was identified as a crucial gap by many researchers.

4. Model: modelling the modified data around the aim of the research. Currently, the scope of the modified data has been narrowed down further to find the contribution of this research. Prominent research is used and simulated to further the understanding of the state-of-the-art methods concerning crowd behaviour analysis. The data directed the aim of research to anomaly detection within crowds. Focus is given to this aim, and further investigating led to the identification of research gaps such as targeting high-density crowds. The contribution of this research has been established throughout this task and documented in Section 1.5.

5. Assess: the assessment of the collected, explored, modified, and modelled data is in an operational state. Experimentation, evaluation, modification, and repetition have been an ongoing process to complete this research's contributions. Assessment is based on a comparison, using evaluation metrics, between the collected data and the results produced by this research.

1.7 Publications

A conference paper titled "Abnormal High-Density Crowd Dataset" was submitted and published by "The Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA2020)". The paper focuses on the novel dataset created as a contribution to this research. The details of the dataset and the application of state-of-the-art crowd anomaly detection methods to this dataset were documented.

The details of the novel approach to crowd abnormality detection using Dynamic Images and CGANs, and the experimentation results are currently being submitted to the IEEE Transactions on Pattern Analysis and Machine Intelligence journal.

1.8 Thesis Structure

The remainder of the research presented in this thesis is organised as follows:

Chapter 2 discusses approaches on how to analyse a crowd stating the most important attributes of a crowd. Notable methods for crowd counting and density estimation are introduced, and a comprehensive discussion of approaches on how to track and re-identify an individual within a crowd.

Chapter 3 discusses noteworthy approaches to crowd behaviour analysis such as recognising and understanding individual/crowd behaviour and detecting anomalies within a crowd. Action recognition methods that are utilised within this research, as well as previous research, are explained. Finally, prominent benchmark datasets and evaluation metrics in the field of crowd anomaly detection are surveyed.

Chapter 4 reviews generative adversarial networks for anomaly detection within a crowd. The applications, types and the basic architecture of GANs are investigated. Focus is given to image-to-image translation using Conditional GANs and the utilisation of this for anomaly detection. Finally, the proposed framework's details combining Dynamic Images and CGANs for crowd anomaly detection are noted.

Chapter 5 describes the high-density crowd dataset created for this research. The data collection process, including resolving privacy issues, pre-processing, and annotating, is described. A description of the dataset is included, and the usage and evaluation methods applicable to this dataset are addressed.

Chapter 6 documents the experimentation and the results achieved by this research. State-of-the-art anomaly detection methods are applied to the high-density dataset created to determine the applicability of these methods on high-density crowds. Dynamic Images merged with CGANs for crowd anomaly detection is applied to benchmark datasets, and the performance results are presented.

Finally, Chapter 7 includes a comprehensive discussion of the results produced and the position of this research in the wider scientific field. Future work and publication of this research are also presented.

2 Background Research

2.1 Introduction

This chapter provides an overview of the significant concepts relevant to the use of computer vision for collective behaviour, and crowd analysis. Collective behaviour is explored to help understand the psychological factors that influence crowds. How and why crowds act in the manner they do has a significant influence on the anylsis of crowds. Subsequently, crowd analysis methods are investigated to determine the conventional aspects considered for analysing a crowd.

2.2 Collective Behaviour

A crowd's hierarchical presentation allows a "crowd" to be ranked at the top level with a collection of multiple groups beneath it (Li et al., 2015). Situated under each group is a collection of individuals; individuals are considered as the bottom level. Mainly a crowded scene can be categorised as either structured or unstructured (Rodriguez et al., 2009). In a structured crowd, the motion of the crowd is usually in a shared direction. The variance in motion direction does not commonly change. Moreover, the crowd exhibits a singular overall behaviour over time. An example of this is footage of an audience in a rock concert; the audience is facing the stage and is most probably swaying together or jumping up and down in unison. As for unstructured crowds (Sjarif et al., 2011), the scene is very hectic (filled with activity, excitement, or confusion); the crowd's motion is random, and the individuals in the crowd move in diverse directions in any given moment. Furthermore, the scene presents numerous crowd behaviours. An example of this is a crowd of commuters in a train station. Individuals in the crowd will exhibit different behaviours such as running to platforms, waiting in line to buy tickets, sitting down on benches, or standing around looking at a screen.

2.3 Crowd Analysis

In this section, multiple factors on how crowd analysis is implemented are investigated. Firstly, techniques on how to recognise a crowd from visual scenes are explored. Secondly, crowd counting/density estimation state-of-the-art algorithms are reviewed. Lastly, current approaches on how to track a person in a crowd throughout multiple images are reviewed.

One definition of a crowd is a collection of individuals in the same physical location, typically with a similar goal shared (Musse and Thalmann, 1997). Crowd analysis is not the same as individual analysis; understanding the crowd's behaviour or an individual in a crowd requires specific approaches specifically adapted for this purpose. An example of an automated crowd analysis framework is shown in Figure 2.

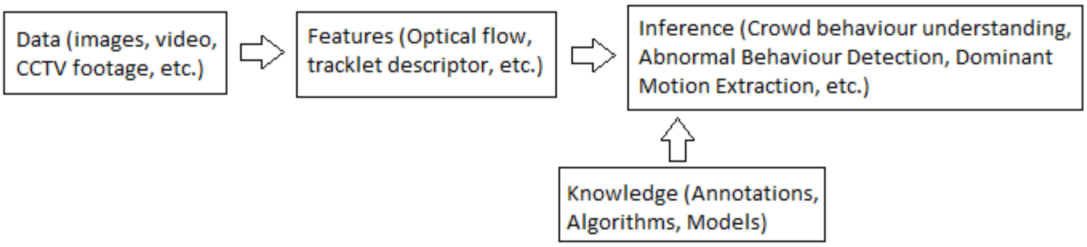

Data (images, video, CCTV footage, etc.)
Features (Optical flow, tracklet descriptor, etc.)
Inference (Crowd behaviour understanding, Abnormal Behaviour Detection, Dominant Motion Extraction, etc.)
Knowledge (Annotations, Algorithms, Models)

Figure 2: Automated crowd scene analysis framework. Adapted from (Li et al., 2015)

An explanation of some of the terms used in crowd analysis (Sjarif et al., 2012; Hasan et al., 2016) is noted in Table 1 below.

Table 1: Terminologies used in crowd analysis.

Crowd Counting / Density Estimation	Measuring a crowd's density status to find the congestion level in an environment or recognise overcrowding.
Crowd Motion Detection	Classifying characteristic of a crowd and extracting crowd behaviour patterns.
Crowd Tracking/ Re-identification	Following a specific person from an image using their trajectories of movement.
Crowd Behavior Recognition	Analysing crowd behaviour to extract temporal information and recognise their behaviour.
Structured Crowded Scene	The crowd's motion is usually in a shared direction. The variance in motion direction does not commonly change. An example of this is footage of an audience in a rock concert; the audience will be facing the stage and sway or jump together.
Unstructured Crowded Scene	The scene is very hectic; the crowd's motion is random, and the individuals in the crowd move in diverse directions at any given moment. The scene presents numerous crowd behaviours; an example of this is a crowd of commuters in a train station.
Pre-Processing	The pre-processing stage includes feature extraction (foreground detection, optical flow), object detection, classification (colour, edge, shape, head, body).
Microscopic	Crowd movement is described as the temporal evolution of each pedestrians' location.
Macroscopic	Crowd movement is described as an averaged spatial representation of individual distribution.
Mesoscopic	Crowd movement is described as a hybrid of Microscopic and Macroscopic
Optical Flow	Displacement or velocity representation of the difference of pixel interval between two consecutive frames.
Tracklet	A tracklet is a fragment of a constructed track following a specific object throughout its movement.

2.3.1 Crowd counting/ Density estimation

Crowd counting and crowd density estimation are central factors in crowd analysis (Sindagi and Patel, 2018). Crowd counting relies on approaches that extract the number of people in a specific scene. In contrast, crowd density estimation is used for more dense crowds to extract the estimated number of people in a scene. This can help solve many issues such as event organisation, public space design, and overcrowding that may lead to stampeding and asphyxia (Grant and Flynn, 2017; Kok et al., 2016). Traditional approaches for crowd counting and density estimation are presented in Appendix B. In this section, however, the state-of-the-art approaches for crowd counting and crowd density estimation are reviewed. A summary of these methods has also been documented in Table 3.

2.3.1.1 State-of-the-art Methods

Liu et al. (2018c) utilise a self-supervised learning approach to crowd counting. This method uses a large number of unlabelled crowd images to enhance accuracy. The idea behind the approach is based on the observation that patches extracted from a high-density crowd ("sub-image") contain a count number equal to or smaller than the "super-image" as shown in Figure 3. The method uses ranked sub-images based on a series of decreasing sized patches to learn the representation of an image. The method achieves an efficient multi-task network that utilises both unlabelled data and the available labelled data to rank the image and estimate the crowd's density maps. Crowd density maps include the spatial distribution information of crowd distribution. Experiments were applied by the authors to two benchmark datasets: UCF Crowd Counting dataset (Idrees et al., 2013) and the ShanghaiTech dataset (Zhang et al., 2016b). The best-achieved results were Mean absolute error (MAE) of 13.7 and Mean squared error (MSE) of 21.4 tested on part B of the ShanghaiTech dataset. The remaining results are documented in Table 2 and compared to other state-of-the-art methods.

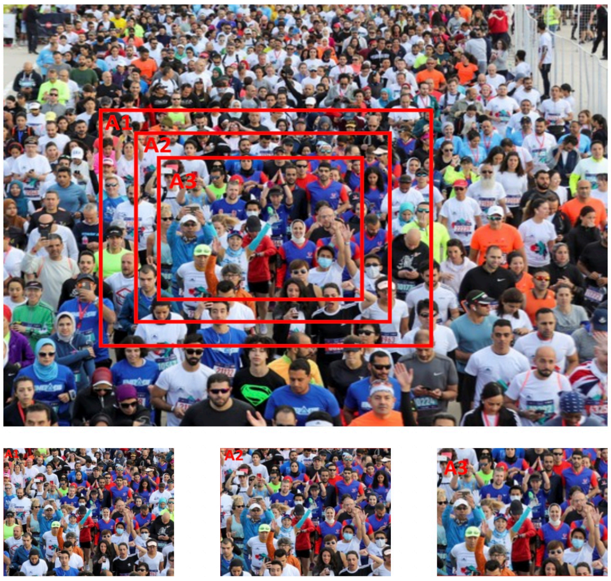


Figure 3: Self-supervised training using sub-image ranking; C(A1) $\geq$ C(A2) $\geq$ C(A3). Adapted from (Liu et al., 2018c)

DecideNet is an end-to-end network presented by Liu et al. (2018a) to estimate a crowd's count. The method extracts two density maps: detection-based map to detect individuals and a regression-based map to extract pixel-wise densities. The maps' variation is based on the motivation that detection-based method estimates more accurately within a low-density crowd but underestimates within a high-density crowd. Furthermore, the regression-based maps overestimate in low-density environments but are more accurate within high-density environments. The authors employ both maps to utilise this fact, and an attention module is used to guide the more fitting estimation based on the crowd's density. However, they found training a fully supervised network is computationally expensive. Testing was applied to three benchmark datasets; their best-compared results were achieved on the Mall dataset (Chen et al., 2012) with 1.52 MAE and 1.90 MSE. The remaining results obtained are noted in Table 2.

The authors of Shen et al. (2018) utilised Generative Adversarial Networks (GANs) to estimate a crowd's count. This research's objectives were twofold. The first is using the GANs for image/patch to generate map translation. A U-net architecture is used to generate the density estimation map from the input patches to achieve this. Meanwhile, adversarial loss is used to weaken the blurriness of the generated density map (shown in Figure 4). Secondly, an adversarial cross-scale consistency pursuit network (ACSCP) was created to conserve the relationship between the whole image input and its patches to ensure the patch crowd count is consistent with the image's overall count. Experiments were applied by the authors to four benchmark datasets. The best results noted were based on testing on the ShanghaiTech dataset part B (Zhang et al., 2016b). The results documented were MAE of 17.2 and MSE of

27.4; other results are shown in Table 2.

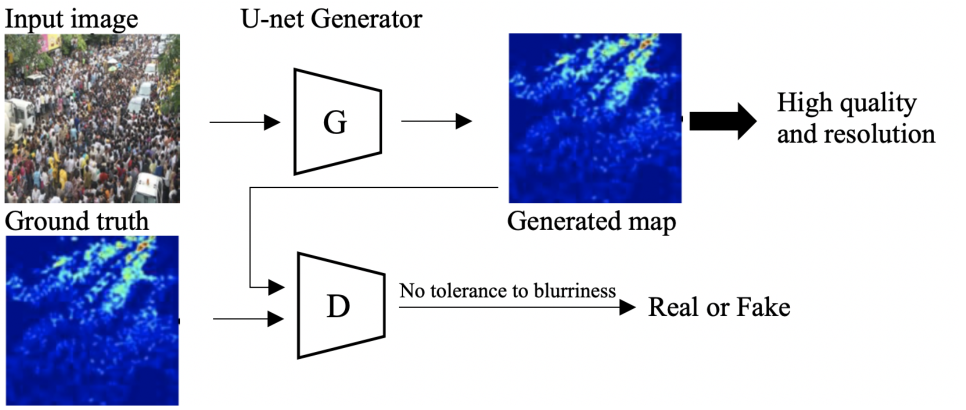


Figure 4: Adversarial loss network for high resolution density map generation. Adapted from (Shen et al., 2018)

Ranjan et al. (2018) presented a multi-branch iterative counting Convolutional Neural Network (ic-CNN). The network was based on two branches and used to generate density maps to estimate a crowd's count from an input image. The first branch of the network was used to produce a low-resolution density map, the produced map and feature maps from this branch are given to the second branch of the network. The second branch of the network utilised the density and feature maps given to estimate the crowd's high-resolution density map. The network architecture is depicted in Figure 5 illustrating the flow of the CNN branches. Experiments were applied by the authors on three benchmark datasets: UCF Crowd counting, ShanghaiTech (Part A and B), and WorldExpo. Their research includes qualitative and quantitative results. The qualitative results are presented in Table 2 to compare with other state-of-the-art methods.

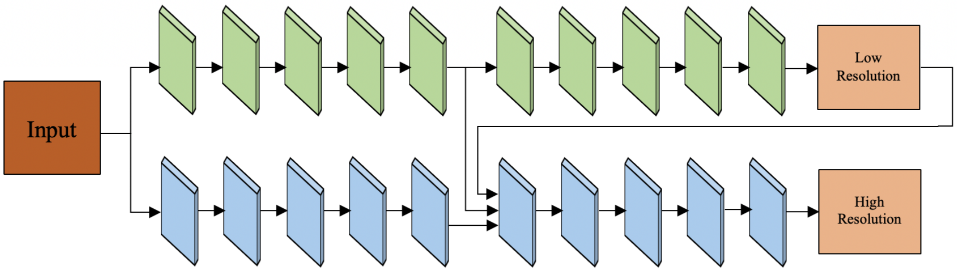


Figure 5: Two-branch iterative counting Convolutional Neural Network. Adapted from (Ranjan et al., 2018)

Jiang et al. (2019) propose a crowd counting and density estimation approach based on a trellis encoder-decoder network named TEDnet. The method uses full images as input and generates density estimation maps of high-quality as output. The network contains several encoder-decoder paths structured hierarchically. The multi-scale encoder can encode the localisation precision to feature maps, which are then used by the multi-path decoder to aggregate and fuse the multi-scale features. The research also better advances the process of backpropagation and the gradient vanishing problem by utilising a blended loss function proposed by the authors. The loss function utilises both spatial abstraction and spatial correlation loss. Testing was applied on four datasets, the best-attained results compared to other methods were achieved using part B of the ShanghaiTech dataset (Zhang et al., 2016b). The results of the MAE were 8.2 and MSE of 12.8; the remaining results are documented in Table 2. Quality of the density maps was also compared to other methods, and this method obtained better results.

Table 2: State-of-the-art Crowd counting/ Density estimation experimental results on various benchmark datasets.

Dataset	UCF CC 50 dataset		ShanghaiTech dataset Part A		ShanghaiTech dataset Part B	
Method	MAE	MSE	MAE	MSE	MAE	MSE
Liu et al. (2018c)	279.6	388.9	73.6	112.0	13.7	21.4
Liu et al. (2018a)	-	-	-	-	21.53	31.98
Shen et al. (2018)	291.0	404.6	75.7	**102.7**	17.2	27.4
Ranjan et al. (2018)	260.9	**320.9**	69.8	117.3	10.4	16.7
Jiang et al. (2019)	**249.4**	354.5	**64.2**	109.1	**8.2**	**12.8**

2.3.1.2 Summary

Crowd counting and density estimation are significant to crowd analysis. The methods discussed have shown variable results when applied. In comparison to other state-of-the-art methods reviewed, the best Mean absolute error (MAE) and Mean squared error (MSE) achieved on various datasets are found in (Jiang et al., 2019). However, there remains a number of major issues that have not been solved as a whole when writing this thesis. The combination of severe occlusion handling, adaptability of static/dynamic movement of people/objects, environmental changes control (weather, background changes, and illumination inconsistency), and real- time implementation, applied efficiently, have not been satisfactorily addressed by the current work. Table 3 summarises the traditional crowd counting and density estimation methods reviewed. The next section discusses another field that is significant to crowd analysis: tracking and person re-identification.

Table 3: Summary of traditional crowd counting/density estimation methods.

Author	Description	Advantages	Approach	Datasets and Results
Cho and Chow (1999); Cho et al. (1999)	The system takes CCTV footage through pre-processing techniques that map the visual data to a two-dimensional feature space using the extracted features. For each image given, three features are extracted: length of crowd edges, the density of crowd objects, and the background density. The neural network takes these extracted feature coefficients as its input. A hybridising of the Least Squares (LS) algorithm and a global optimisation method is used as the system's learning algorithm to classify the crowd.	Neural-based system.	Pixel-based, indirect approach	Own dataset. 1.189 SSE using the LS and SA algorithm and estimation accuracy of 94.36%.
Cho et al. (1999)	The previous authors enhance their system with a cross-over between the LS algorithm and Genetic Algorithm (GA). The same neural network topology as previous is used.	Neural-based system.	Pixel-based, indirect approach	Own dataset. Estimation accuracy result of 93.8%, but with decreased CPU running time.
Lin et al. (2001)	A single image is used to make an approximation of the number of people in a crowd. The system extracts the contour of people's heads using the Haar Wavelet Transform (HWT) function (Chapelle et al., 1999). A support vector machine (SVM) will determine if the extracted data is a "head" or not. Perspective transformation (also known as imaging transformation) is used for crowd size estimation.	The algorithm can handle background scenes of a complicated nature.	Model-based, direct approach	Own simulated world. Accuracy level of 90% - 95%.
Viola et al. (2003)	The authors take advantage of both image appearance data and motion data by fusing them together. The image appearance data used is based on feature extraction using an integral image. The detector is then trained using AdaBoost (Schapire and Singer, 1999).	Applicable to low-resolution images and bad weather conditions.	Model-based, direct approach	Own dataset. 1 in 400,000 false positives, 80% detection rate.
Brostow and Cipolla (2006)	The algorithm is based on an unsupervised Bayesian clustering algorithm. The approach assumes that a pair of points, that are parallel in movement, probably belong to the same object. Low-level features are extracted and clustered using probabilistic behaviour. The paper considers tracing two features: both Rosten-Drummond features (Rosten and Drummond, 2006) and Tomasi-Kanade features (Tomasi and Detection, 1991). Additionally, to track the features in two frames, hierarchical optical flow is applied.	No training required.	Trajectory-based, direct approach	Own dataset.94% detection rate, 22.9% false detection rate.
Kong et al. (2006)	The system uses a feed-forward neural network in a viewpoint invariant learning-based method to map the connection between the number of pedestrians and the feature histogram extracted from low-level features.	Use of feature histograms instead of simple features, the system uses a supervised feed-forward neural network.	Corner Point-based, indirect approach	Own dataset. Results presented in graphical format.
Sidla et al. (2006)	The system finds Ω-like shape information for human detection and implements texture feature for human recognition. A description of each person is deduced from the use of a co-occurrence matrix feature vector by using the Kalman filter as well as, Kanade-Lucas-Tomasi (KLT) (Tomasi and Detection, 1991) tracking points are used.	Results are similar in indoor and outdoor scenes.	Trajectory-based, direct approach	Own dataset. Absolute mean error 2%-10%.

Table 3: Summary of traditional crowd counting/density estimation methods.

Author	**Description**	**Advantages**	**Approach**	**Datasets and Results**
Chan et al. (2008)	The system used a combination of dynamic textures to segment the crowd into multiple motion features. The method uses a Gaussian process for counting the number of people.	Estimation without tracking or explicit object segmentation.	Texture-based, indirect approach	UCSD dataset. MSE of 4.181 (away) and 1.29 (towards).
Zhao et al. (2008)	A stochastic approach within the Bayesian framework that can model multiple partially occluded humans. The method begins by applying blob boundary detection, then canny edge detection and then the head and shoulder model is applied. Edge intensity is used for reliable detection. The data-driven Markov chain Monte Carlo (DDMCMC) sampling configuration is used for foreground estimation.	Humans do not have to be un-occluded when they first enter the scene.	Model-based, direct approach	Own dataset and CAVIAR. Results on own dataset: 98.13% detection rate, 0.27% false detection rate.
Chan and Vasconcelos (2009)	Continuation to the previous method; the system is based on a standard Poisson regression model in a Bayesian setting. The developed predictive distribution was kernelised, and representations of non-linear log-mean functions were admissible. Estimated marginal likelihood function was developed and used to show its relation to a Gaussian process with a special non-i.i.d. noise term.	Estimation without tracking or explicit object segmentation.	Texture-based, indirect approach	UCSD dataset. MSE of 2.4675 (away) and 2.0246 (towards).
Ryan et al. (2009)	Using a foreground subtraction technique; the local features are extracted with respect to blob segments. The number of people in each blob segment is estimated so that the accumulation of all the segments in the scene is the scene estimation.	Extraction of local features instead of holistic features.	Texture-based, indirect approach	UCSD dataset. MSE of 3.065.
Chen et al. (2012)	An automated model that learns the functional mapping between multi-dimensional structured output and interdependent low-level features. The model can count people by finding the intrinsic importance of multiple features.	Can handle diverse environments.	Texture-based, indirect approach	UCSD dataset, own Mall dataset. MSE of 8.08 (UCSD) and 15.7 (Mall dataset).
Liang et al. (2014)	The system uses an altered more robust SURF (Speeded Up Robust Feature) algorithm. Additionally, an enhancement to the DBSCAN (Density-Based Spatial clustering of Application with Noise) (Ester et al., 1996) clustering algorithm is applied. The system tracks feature points by using the combination of local Lucas-Kanade optical flow (Lucas and Kanade, 1981) and Hessian matrix algorithm to determine the crowd flow orientation.	The system can also detect crowd flow orientation.	Corner Point-based, indirect approach	PETS dataset. MAE for four different video sequences of 1.01%, 1.17%, 4.33%, and 1.39% respectively.
Tang et al. (2015)	The system uses several cameras for diverse views to gather corresponding data. Crowd count and normalising the visual feature perspective are considered as one learning problem. The algorithm receives multiple views of a crowd and matches them. Regressors count the crowd using intra-camera images and inter-camera predictions conflict.	Multiple crowd views.	Pixel-based, indirect approach	PETS 2009 dataset. 3.26(FPR), 2.52(TPR).
Liu et al. (2018c)	Learn the image representation by using ranked sub-images based on a series of decreasing sized patches. By using available labelled data to rank image and estimate the density maps of a crowd.	Makes use of unlabelled images to enhance accuracy.	Self-supervised approach	UCF CC and ShanghaiTech dataset. MAE of 13.7, MSE of 21.4 on ShanghaiTech Part B.
Liu et al. (2018a)	DecideNet is an end-to-end trained network that estimates the count of a crowd by extracting two density maps: detection-based map to detect individuals (better in high-density) and a regression-based map to extract pixel-wise densities (better in low-density).	Utilise both detection and regression-based maps.	End-to-end trained network	ShanghaiTech Part B and Mall dataset. Results of 1.52 MAE and 1.90 MSE.

Table 3: Summary of traditional crowd counting/density estimation methods.

Author	**Description**	**Advantages**	**Approach**	**Datasets and Results**
Shen et al. (2018)	Utilise GANs for image-to-patch translation and back. An ACSCP network conserves the relationship between the whole image and extracted patches to ensure consistency between patch count and overall count.	A novel approach in utilising GANs for crowd counting.	Generative Adversarial Networks	UCF CC and ShanghaiTech A & B. ShanghaiTech dataset part B results are MAE of 17.2 and MSE of 27.4.
Ranjan et al. (2018)	A multi-branch CNN combines both low-resolution density map and feature maps (first branch) to produce high-resolution density map.	Produces a high-resolution density map of the crowd.	Convolutional Neural Network	UCF CC, ShanghaiTech and WorldExpo. MAE and MSE of 260.9, 320.9 & 69.8, 117.3 & 10.4, 16.7 respectively.
Jiang et al. (2019)	TEDnet contains several encoder-decoder paths structured hierarchically. The encoder can encode the localisation precision to feature maps used by the decoder to aggregate and fuse the multi-scale features.	Advances gradient vanishing problem by utilising a blended loss function.	Trellis encoder-decoder network	UCF CC, ShanghaiTech and WorldExpo. ShanghaiTech Part B MAE 8.2 and MSE of 12.8.

2.3.2 Tracking / Person Re-Identification

The definition of person re-identification is the following of a specific person from an image taken from one camera and re-identifying them in an image from a different camera (Lavi et al., 2018). This is a challenging field compared to normal tracking algorithms; there are many more issues to consider. Some of the problems that are associated with tracking/person re-identification are the ambiguity in visuals and the uncertainty of the spatial and temporal human appearance across various cameras (Kasturi and Ekambaram, 2014). Contextual and non-contextual methods for tracking and person re-identification are discussed in Appendix C and the state-of-the-art methods are reviewed below. A summary of these methods has also been documented in Table 4.

2.3.2.1 State-of-the-art Methods

The authors of Ristani et al. (2016) apply a multi-target and multi-camera (MTMC) tracking system originally applied for multi-target single-camera tracking. The method combines target detections received from a detection system into tracklets. An easy motion model is utilised in this method as the aggregated tracklets are short enough to model. The method generates identities; which are single-camera trajectories (combined tracklets) connected to multi-camera trajectories. The authors also propose evaluation metrics to identify how frequently a target identified accurately: Identification Precision (IDP), Identification Recall (IDR), and F1 score (IDF1). These evaluation metrics were used to test their method on a benchmark dataset and their own dataset (DukeMTMC). The Upper bound results achieved on various cameras from the DukeMTMC dataset were 72.25 IDP, 50.96 IDR and 59.77 IDF1.

An open-world person re-identification system applied to dense crowds is presented in Assari et al. (2016). The method combines several Personal, Social and Environmental (PSE) restraints to model human motion throughout cameras with high-density crowds. The PSE restraints are preferred speed, destination, spatial grouping and social grouping. The preferred speed is an assumption of the walking speed of the persons in a crowd. The destination restraint is the destination probability of an individual calculated by observing recurring motion patterns. Finally, the spatial and social grouping is calculated based on the span persons have travelled from one point to another and a reward system of persons that travel together in groups. Unlike the previous state-of-the-art methods documented, this method applies experimentation on Grand Central Station dataset (Zhou et al., 2012), with 97.31% Area Under Curve (AUC), and 84.19% F-score. It is one of the first methods to combine all three PSE constraints in modelling human motion.

Spindle Net, (Zhao et al., 2017), is a Convolutional Neural Network (CNN) built on human body structure data for representation learning. The network extracts semantic body features from several regions of the body to be matched throughout images. During the stages of feature extraction some of the features are maintained, they are then mingled in a fusion network. The network helps extract discriminative features of persons and match regions of the body across images. This network is one of the first to utilise body structure information for re-identification across different cameras (images). Experimentation was applied on seven datasets for re-identification (ReID). However, these datasets do not intersect with the research reviewed in this research to compare. Compared to other methods mentioned by the authors, their method outperforms them regarding Top-1 accuracy.

A pose normalised generative adversarial network (PN-GAN) designed by Qian et al. (2018) is used to re-identify individuals throughout multiple cameras. The deep model is used to reduce the impact of large pose variations. As shown in Figure 6, the framework begins by using an input image with an individual with an initial pose and generates a synthesised image of the individual with an intended pose (pose-normalised image). Two sets of features are extracted from the ReID model after it is trained on the original image and the synthesised image. The features are combined to create an output descriptor. The model is adaptable to new re-id datasets without fine-tuning the model to the new training data. The evaluation metrics used in this research are rank-1 (R-1), rank-5 (R-5), rank-10 (R-10) accuracy, and mean average precision (mAP). The model has been tested on multiple datasets, and results of R-1, R-10 and mAP on the DukeMTMC (Ristani et al., 2016) dataset are 73.58, 88.75 and 53.20 respectively.

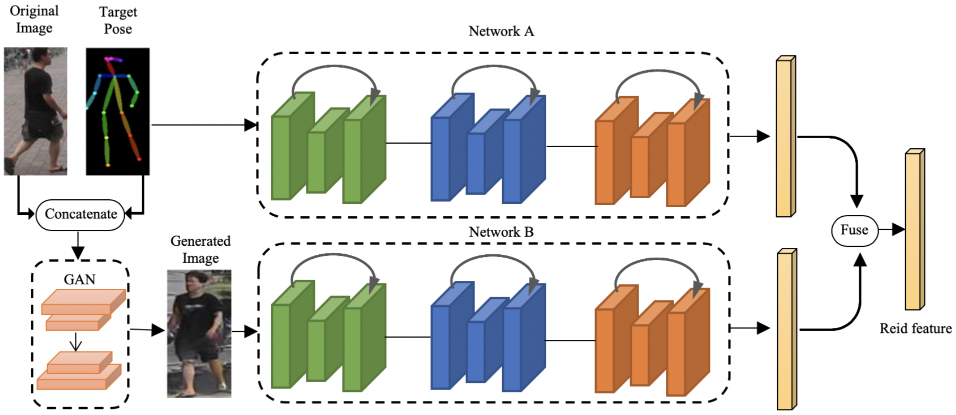


Figure 6: Pose-Normalised Generative Adversarial Network (PN-GAN) framework. Adapted from (Qian et al., 2018)

Ristani and Tomasi (2018) design a convolutional neural network (CNN) that utilises both Multi-Target Multi-Camera Tracking (MTMCT) and Person Re-Identification (Re-ID) features for MTMCT purposes. MTMCT is used to track multiple people through multiple cameras, whereas Re-ID can identify a targeted person in multiple images. The framework shown in Figure 7 starts with extracting bounding boxes of detected individuals, then motion and appearance features are extracted to infer trajectories. Correlation clustering optimisation is then used to deduce and label correlations based on the extracted trajectories. Lastly, missed detections are introduced, and low confidence trajectories are removed. Testing was applied to multiple benchmark datasets, and evaluation metrics used were IDP, IDR, IDF1 and multiple Object Tracking Accuracy (MOTA). The results noted on the DukeMTMC dataset with the best detector and feature configurations were 83.50 IDP, 77.25 IDR and 80.26 IDF1.

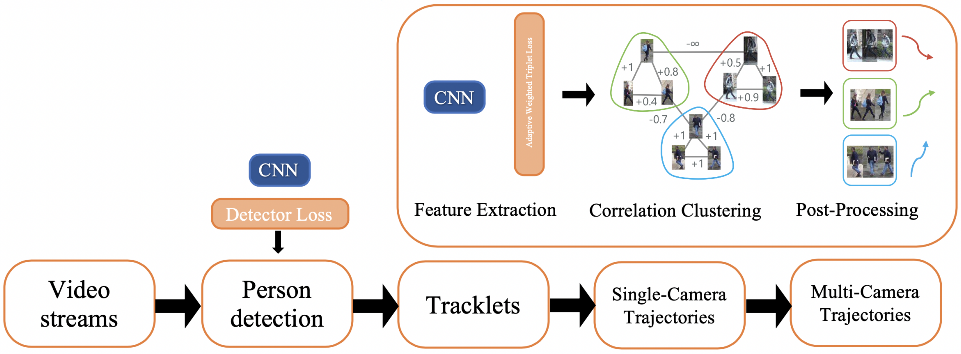


Figure 7: Multi-Target Multi-Camera Tracking (MTMCT) and Person Re-Identification (Re-ID) framework. Adapted from (Ristani and Tomasi, 2018)

2.3.2.2 Summary

Table 4 summarises the tracking/person re-identification methods reviewed. The previously noted work is noticeably applicable to specific real-world problems and progressions of other fields can help improve methods in this area. For instance, there have been great advancements in biometric data extraction from lower quality footage. This biometric data can be incorporated into person re-identification algorithms (Tavares et al., 2019). Moreover, integrating representational models of the associations between low-level features and high-level semantics could improve the scalability and computational complexity issues presented within this field.

Table 4: Summary of traditional Tracking/Person Re-Identification methods.

Author	Description	Advantages	Approach	Datasets and Results
Gandhi and Trivedi (2007)	The system uses a Panoramic Appearance Map (PAM) to extract features of a targeted object from footage taken from multiple cameras. Features are combined to create a single signature, to find the position of the object multiple-camera triangulation is used to place a cylinder-shaped model around the object.	For the approach to work properly there is a requirement that three or more cameras simultaneously view the object.	Contextual Approach.	Own dataset. PAM results are presented in graphical format.
Javed et al. (2008)	The algorithm learns the space-time cues and therefore learns the inter-camera connection; which are used to constrain a relationship between cameras. With the use of kernel density estimation, the relationships are modelled as probability density functions of space-time variables such as entrance/exit locations, velocity, and transition times.	Applies brightness transfer function to handle appearance alterations between cameras. The system learns camera topology.	Contextual Approach.	Own dataset. Tracking accuracy results are presented in graphical format.
Bazzani et al. (2010)	Identification signature named Histogram Plus Epitome (HPE); features extracted from multiple images of a human are concentrated to develop the signature. Redundant or outlier images are removed using unsupervised Gaussian clustering technique. The human appearance is then described using two complementary features: global and local. Appearance matching is implemented through a weighted sum of feature similarities.	Occlusions and crowded scenes are handled well.	Non-Contextual Approach.	i-LIDS and ETHZ datasets. Comparative results are noted in graphical form.

Table 4: Summary of traditional Tracking/Person Re-Identification methods.

Author	Description	Advantages	Approach	Datasets and Results
Baltieri et al. (2011)	The approach is built on three key modules: detection of an object, short-term tracking, and long-term tracking. Detection uses merged information extracted from all camera views to find object location on the ground plane. For short-term tracking, the Kalman filter is used to track individuals, and Local matching is achieved using geometrical and spatial data. Long-term tracking finds the trajectories corresponding to the same object and then matches and combines them for re-identification.	The system has short-term tracking and long-term tracking.	Contextual Approach.	PETS and EPFL Terrace indoor datasets. TER (Total error rate) of 10% (PETS) and 7% (EPFL), long-term tracking precision results of 72.73% and recall results of and 88.8%.
Bazzani et al. (2013); Farenzena et al. (2010)	The descriptor is constructed with symmetry-driven appearance-based features combined with a simple distance minimisation technique for object matching. The system localises body parts and removes unnecessary background data. The localised parts are used to extract three corresponding appearance characteristics necessary for both re-identification and multiple target tracking.	The descriptor can handle pose, viewpoint, and illumination changes.	Non- Contextual Approach.	CAVIAR. False positives, false negatives, and average tracking accuracy of 0.0608, 0.1852, and 0.4567, respectively.
Ristani et al. (2016)	Target detections received from a detection system into tracklets. Then identities are created from single-camera trajectories (combined tracklets) and connected to multi-camera trajectories.	Multi-target and multi-camera (MTMC) tracking.	Motion model from aggregated tracklets.	DukeMTMC dataset were 72.25 IDP, 50.96 IDR and 59.77 IDF1.
Assari et al. (2016)	Personal, Social and Environmental (PSE) constraints are used to model human motion throughout cameras with high-density crowds. The PSE restraints are preferred speed, destination, Spatial Grouping and Social Grouping.	A novel approach utilising PSE restraints.	Open-world person re-identification system using PSE.	Grand Central Station dataset with 97.31% AUC and 84.19% F-score.

Table 4: Summary of traditional Tracking/Person Re-Identification methods.

Author	**Description**	**Advantages**	**Approach**	**Datasets and Results**
Zhao et al. (2017)	Extracts semantic body features from several regions of the body as discriminative features to be matched throughout images. During feature extraction, some features are maintained and mingled in a fusion network.	A novel method utilising body structure for re-identification.	Convolutional Neural Network.	Seven datasets for ReID, high Top-1 accuracy.
Qian et al. (2018)	Two sets of features are extracted from the trained model on the original image and the synthesised image. The features are combined to create an output descriptor to re-identify persons.	Model is adaptable to new re-id datasets without fine-tuning.	Pose normalised generative adversarial network.	DukeMTMC dataset: R-1, R-10 and mAP of 73.58, 88.75 and 53.20 respectively.
Ristani and Tomasi (2018)	Extracts bounding boxes for individuals then motion and appearance features are extracted to infer trajectories. Correlation clustering optimisation is used to deduce and label correlations based on the extracted trajectories.	Utilises both Multi-Target Multi-Camera Tracking (MTMCT) and Person Re-Identification (Re-ID).	Convolutional Neural Network.	DukeMTMC dataset: 83.50 IDP, 77.25 IDR and 80.26 IDF1.

2.4 Conclusion

As the focus of this research is crowd anomaly detection using computer vision techniques, a comprehensive overview of collective behaviour and crowd analysis have been presented. The fundamental psychological factors behind collective behaviour assisted in the understanding of crowd influences. The detection and recognition of collective acts has a significant impact on crowd analysis. More specific fields such as crowd counting/density estimation and tracking/person re-identification have been covered to further understand the standard and state-of-the-art methods used to analyse a crowd. The next chapter (Chapter 3), continuing the literature review, investigates a more specific field "Crowd Behaviour Analysis" including the main focus of this research: crowd anomaly detection.

3 Crowd Behaviour Analysis

3.1 Introduction

This chapter provides a comprehensive review of the field of crowd behaviour analysis. This field is explored to help understand anomaly detection requirements within crowds, which is one of the main focuses of this research. Initially, strong handcrafted techniques for motion representation and detection of anomalous behaviour are discussed. Afterwards, state-of-the-art neural networks methods are explored for the same purposes. Finally, the most recent and novel approaches to crowd behaviour analysis using generative adversarial networks (GANs) are examined. For the purposes of this research, the methods and technologies that are examined are directly related to crowds. Notable reviews and surveys, published by various researchers, in crowd analysis and crowd behaviour analysis, were reviewed and documented in Appendix A. These reviews helped provide a view of the existing work in crowd analysis in an organised and comprehensive way.

3.2 Motion Representation

Crowd motion representations are specific crowd features extracted for the purpose of analysing crowd behaviour. Crowd behaviour analysis methods extract various types of motion representations for the detection and/or identification of crowd behaviour. Noteworthy and novel methods in this field are investigated below.

A notable framework proposed by Ali and Shah (2007) is applied to high-density crowds for segmentation and flow instability detection purposes. The Lagrangian Particle Dynamics structure is used for particle advection based on the flow fields generated by the moving crowd. The authors handled moving crowds as an aperiodic dynamical system. Advection is the process of matter moving along or becoming advected by a flow. These flows can be modelled using a velocity field; specifying the velocity at a specific position and time. 'Flow segments' were used as an indication of the emerging motion patterns. The authors presented a flow segmentation structure based on non-linear dynamical systems, fluid dynamics, and turbulence theory to find these flow segments. The trajectories extracted from the particle advection would shine a light on essential flow features, which have a direct correlation with physical objects within a scene. As for flow instability detection, the authors consider any change of flow segments to be abnormal. A connection between flow segments over time is created, and the occurrence of a new flow segment indicates normal flow abnormality. Testing the approach was applied on high-density crowd/traffic scene videos taken from the stock footage web sites such as Getty-Images, Photo-Search and Video Google. Additionally, video footage from a National

Geographic documentary named 'Inside Mecca' is used to experiment further. Although quantitative results are not documented, qualitative results for both flow segmentation and detection of flow instability are noted when tested using the footage mentioned above.

Alternatively, (Wang et al., 2007) propose an unsupervised learning framework that uses data extractions from visual material to understand actions within crowded scenes. Their Bayesian model is used to link low-level visual data, "atomic" simple actions, and multi-agent connections. The atomic actions are modelled using the extracted low-level visual data; furthermore, the multi-agent connections are modelled using the atomic actions. This framework does not track humans but instead uses local motion for features. System performance deteriorates due to the partitioning of extended footage to shorter, and more manageable clips. The authors do not note any quantitative evaluation metrics, but testing was applied to a 1.5 hour-long traffic scene dataset, and the results are documented as figures.

Cheriyadat and Radke (2008) describe a method for the identification of dominant motions within a crowd. They use an optical flow (further explained in Section 3.4.1) algorithm to track low-level object features. More specifically, they use the Shi-Tomasi-Kanade (Shi and Tomasi, 1994) and the Rosten-Drummond (Tomasi and Detection, 1991) detectors to extract the low-level features. Then, an upgraded implementation of the Kanade-Lucas-Tomasi optical flow algorithm (Lucas and Kanade, 1981) was used to track these features. As a result, feature point tracks are extracted, but they were long and considered undependable leading to the need for a clustering method. The longest common subsequence was used as a distance metric to compare feature point tracks. The tracks that were alike in direction and considered as spatially nearby were clustered together having an outcome of smooth dominant motions. Quantitative metrics were not applied, but experiments were applied on four different video footage sequences: Platform sequence, Campus sequence, Escalator sequence, and Airport sequence taken from the PETS 2007 benchmark dataset (Ferryman and Tweed, 2007). In video format, the authors' document, the feature points, the point tracks, and the dominant motions for each video sequence in their research (Cheriyadat and Radke, 2008).

Curl, and Divergence of motion Trajectories descriptor (CDT) for behaviour analysis is presented by Wu et al. (2017). The descriptors are found using curl and divergence along tangential and radial paths that denote trajectory motions and their respective conjugate fields. In addition to using the CDT to describe the collective motion sequence, the method considers both local characteristics and global structure of a motion vector field. Finally, to classify the crowds' behaviour, the authors initially extract sub-motion fields from the motion vector fields using particle advection. The method is robust to overlapping motion patterns and can discriminate amongst them. The authors then employ max-min pooling and dense motion to excerpt a cohesive feature vector of rich motion data. For experimentation, the CDT descriptors are limited to five identifying behaviours; lane, clockwise arch, counter-clockwise arch, bottleneck and fountain-head. The proposed method is compared to four other methods and tested on the UCF (Idrees et al., 2013) and CUHK (Shao et al., 2014, 2017) datasets. Thoroughly presented are the results of various testing setups. Moreover, the quantitative results such as ROC, true-positive and false-positive rates, and experimental graphs show favourable results from this technique.

3.3 Crowd Anomaly detection

Generally defined, anomalies within a crowd are atypical patterns that do not conform with the learnt normality (Singh et al., 2020). Anomaly detection is also typically considered an outlier detection problem where an abnormality would be a low-probability event regarding a learnt normal behaviour model (Mahadevan et al., 2010). Crowd anomaly detection is applied to detect anomalous or non-typical scenes within footage of a crowd. This application is essential in the prevention of crowd disasters in fields such as video surveillance. There are two main methods predominantly used in crowd anomaly detection: hand-crafted methods and machine learning methods:

- **Handcrafted methods:**
  These methods require the extraction of motion and/or appearance features such as optical flow and tracklets. Traditionally, to reconstruct normal scenes with small reconstruction errors, a taught dictionary is used. On the other hand, the features that match to anomalous scenes would have large reconstruction errors. The problem with this method is that it requires incorporating some priori knowledge during training. This incorporation can be complicated in cases of complex video surveillance scenes.

- **Machine Learning:**
  Some of the supervised and unsupervised methods are convolutional neural networks (Sabokrou et al., 2018), convolutional auto-encoders (Fan et al., 2020), stacked denoising auto-encoders (Vu et al., 2019), spatio-temporal auto-encoders (Fradi et al., 2017), and long-short term memory (Majumder et al., 2018). They tend to do better with unsupervised methods than supervised ones due to the scarcity of annotations and small training data size. These methods usually incorporate low-level features such as lines, curves and edges, or high-level features such as object and shapes. The problems with using just low level-feature detection are:

  - It usually causes fragmented and interrupted regions; and
  - It is sensitive to noise and is significantly affected by environmental changes.

- Deep machine learning methods that incorporate **Generative Adversarial Networks (GANs)** in their framework have presented accuracy results that surpass other deep learning models. The conditional GANs (CGANs) are trained to translate between a pair of frames and their corresponding optical flow features using image-to-image translation (Isola et al., 2017). The CGANs are then used to generate either frames or optical flow based on the input. CGANs have previously been incorporated with CNNs, autoencoders and denoising autoencoders for crowd anomaly detection.

An overall summary of the general computer vision methods of crowd behaviour analysis and crowd anomaly detection in video monitoring is illustrated in Figure 8.

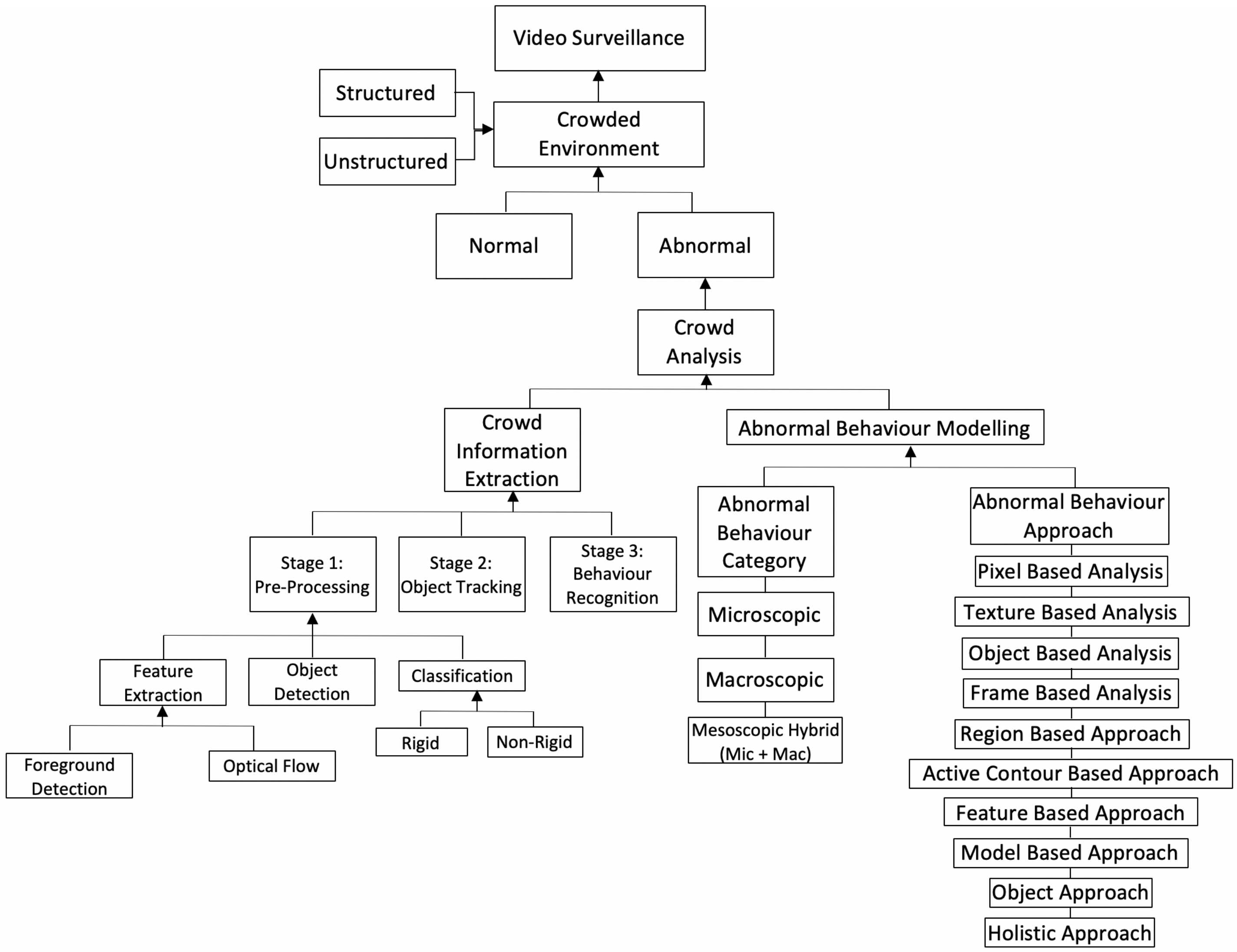


Figure 8: Crowd abnormal behaviour detection framework in video monitoring. Adapted from (Sjarif et al., 2012)

3.3.1 Criteria for anomaly detection in images

There are two standard criteria used to consider an anomaly (any behaviour outlying from established normalcy) within an image: frame-level and pixel-level abnormality detection (Li et al., 2014). The third criteria is a dual pixel-level (Sabokrou et al., 2015), which considers the pixel-level constraint as well. These constraints are used to calculate the true-positive rate (TPR) and the false-positive rate (FPR) (further explained in Section 3.5). Explained below are each of the constraints:

- **Frame-level detection:** This criterion of detection does not consider the localisation of anomalies within a frame. Instead, if any pixel within the frame is detected as abnormal, the whole frame is considered abnormal. If the ground truth data coincides with the frame detection, a true-positive is tallied up into the TPR. To compute the ROC curve (Section 3.5) this detection method is applied several times using different thresholds. (Mahadevan et al., 2010; Li et al., 2014; Ravanbakhsh et al., 2017)
- **Pixel-level detection:** This criterion of detection considers the importance of abnormality localisation within a frame. The requirement is; at least 40% of the anomaly ground truth pixels are covered by the detected pixels. This detection's weakness is "Lucky Guess"; if a part of the detected region overlaps with the ground-truth data, the false detected regions are not taken into consideration. An additional criterion (Dual pixel-level), is used to solve this. (Mahadevan et al., 2010; Li et al., 2014; Ravanbakhsh et al., 2017)
- **Dual pixel-level detection:** This novel criterion of detection applies the pixel-level detection constraint and requires that at least β% of the detected pixels are covered by the anomaly ground truth pixels. (Sabokrou et al., 2015; Vu et al., 2019)

3.3.2 Anomaly detection using Handcrafted methods

This section presents a comprehensive investigation of handcrafted, neural network and generative adversarial network methods for crowd anomaly detection.

Andrade et al. (2006) model the normal behaviour of a crowd using an unsupervised feature extraction method. The extraction method fits an HMM (Hidden Markov Model) for all the footage fragments; afterwards, spectral clustering is applied using a calculated similarity matrix. Using the clustered fragments, the authors discover the appropriate number of models to characterise normal motion patterns by training a new set of HMMs. New footage is compared to the normal behaviour models using a detection threshold to detect anomalous crowd behaviour. Two simulated datasets are used by the authors for experimentation, one with normal crowd flow and the other with footage of a congested exit. Quantitative measurements are not documented, but a visual representation of the likelihood function results demonstrated the approach's effectiveness.

A social force model (SFM) is used by Mehran et al. (2009) for the detection and localisation of abnormal crowd behaviour. To achieve this, the authors used particle advection founded on the space-time average of optical flow (further explained in Section 3.4.1). The method considered the moving particles to correspond to the individuals. Moreover, the SFM is used to estimate the interaction force between them. Bag of words approach uses a vector field, the mapping of interaction forces to image frames, to model the crowd's "normal" behaviour. The UMN dataset (University of Minnesota, 2006) and a web dataset of footage gathered from Getty Images, and ThoughtEquity.com were used in experimentation. The authors report results from the UMN dataset (noted in Table 7) testing with 0.96 area under ROC. This is an enhancement compared to the pure optical flow for detecting abnormal crowd behaviour method, which demonstrated a result of 0.84 area under ROC. UCSD results are noted in Table 5.

Using interaction energy potentials, Cui et al. (2011) presented an approach linking the existing state of a subject's behaviour and its corresponding action. The existing state and the subjects' actions are represented by the interaction energy potential function and velocity. For crowd interaction modelling, spatio-temporal points of interest are extracted and tracked, eliminating the need for humans' recognition and segmentation. Additionally, this makes the method more robust to errors that arise with recognition and segmentation techniques. Finally, with the use of an SVM, an anomaly can be detected when the extracted Energy-Action forms seem unfamiliar, a representation of the framework is shown in Figure 9. Experimentation is applied to the BEHAVE (Blunsden and Fisher, 2010), and the UMN datasets (University of Minnesota, 2006) and results are documented in figure format. The method is valid on reasonably crowded scenes and shows improved results compared to pure optical flow and SFM (Mehran et al., 2009) previously mentioned in this section.

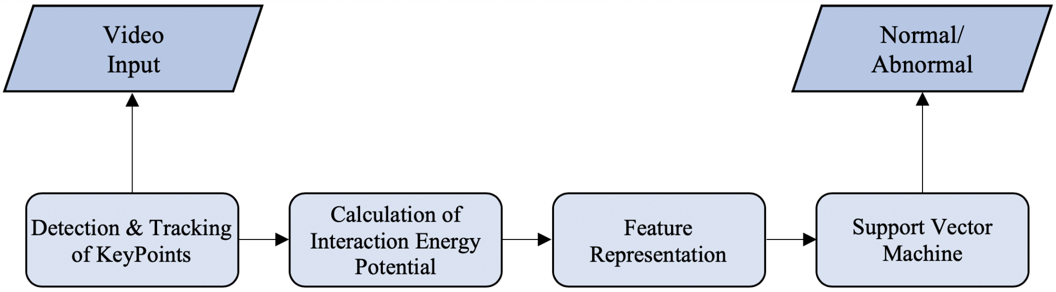

Video Input
Normal/ Abnormal
Detection & Tracking of KeyPoints
Calculation of Interaction Energy Potential
Feature Representation
Support Vector Machine

Figure 9: Flow chart of Interaction Energy Potentials framework. Adapted from (Cui et al., 2011)

Based on the typical social force model, Yang et al. (2012) present a local pressure model that considers the crowd's local characteristics. The method can detect an anomaly within a crowd using local velocity and local density characteristics. The method, shown in Figure 10, starts with the placement of a grid of particles to calculate the local characteristics efficiently. A pressure model is used to extract local pressure using these characteristics. Consequently, feature vectors are extracted for the footage frame with the utilisation of Histogram of Oriented Pressure (HOP). Finally, for abnormality detection, a Support Vector Machine (SVM) is used for classification, and a median filter is implemented on the classification results for further improvement. A median filter is a non-linear digital filtering method typically used to remove noise from a signal or image. The algorithm typically runs through an image and substitutes each entry with the median of the neighbouring entries. Experimentation is applied on the UMN dataset (University of Minnesota, 2006), and an area under curve (AUC) value of 0.9784 is noted as a better value in comparison to the SFM method (Mehran et al., 2009).

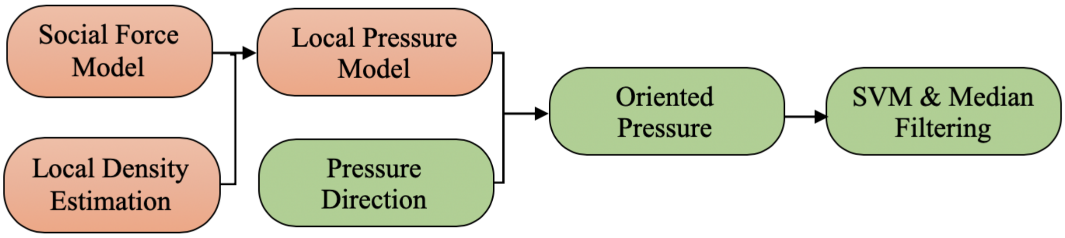


Figure 10: Structure of Anomaly Detection System. Adapted from (Yang et al., 2012)

Using sparse combination learning and bag-of-words, (Lu et al., 2013) can detect abnormal events for robust inference. The framework resizes frames into several scales to partition layers in a uniform manner to create a set of non-overlapping patches. A spatial-temporal cube is created using this data and used to extract 3D gradient features. Based on spatial matching, the extracted features are independently processed for training and testing. Fundamentally, the research assumes the beginning part of the input video contains normal behaviour, so the normal behaviour dictionary created using it. In testing, reconstructing abnormal behaviour from the normal behaviour dictionary presents high reconstruction errors; this is how an abnormal event is detected. Additionally, the method can process an average of 140Ĩ50 frames per second. This method presents high false alarm rates due to the variety of environmental changes throughout a video sequence. Experimentation was applied on their dataset (Avenue dataset), Subway dataset and UCSD Ped-1 Dataset. The results of UCSD and Avenue are noted in Tables 5 and 6.

An innovative spatio-temporal method for crowd modelling is presented by Fradi et al. (2017), the authors use the model to extract visual descriptors representing the crowd. Initially, for crowd representation, the method uses a Delaunay graph over time for dynamic emulation. Moreover, a spatial feature is integrated into the graph for overall completeness to extract the interactive descriptors. Using both the spatial and temporal information, the authors claim to define a novel set of visual descriptors. Consideration is given to interactive and distinct behaviours within the descriptors, and abundant semantic crowd data is encoded. The use of the Delaunay graph is the primary provider to this method, and the extracted local entities data joined with tracklet data is novel in being applied to crowd analysis. Three applications are considered in experimentation; crowd video classification, crowd anomaly detection and localisation, and crowd violence detection. The CUHK (Shao et al., 2017, 2014) for crowd classification dataset was used in the first experiment. The accuracy results equate to 85.25% while noting comparisons to other techniques this method presents better accuracy results. The second experiment for crowd anomaly detection and localisation used the UMN dataset (University of Minnesota, 2006) and presented an average result (over three scenes) of 98.61 area under the curve (AUC). Also, documented were comparison results with six other methods, four of them that proved better results. Lastly, an experiment for crowd violence detection on the violent-flows dataset (Hassner et al., 2012) was applied, this exhibited much better results compared to other methods with 84.44% accuracy and 88.00 AUC.

El-Etriby et al. (2017) examine an innovative framework utilising discriminative models such as Conditional Random Field (CRFs), Hidden Conditional Random Field (HCRFs) and latent dynamic Conditional Random Fields (LDCRFs) to detect crowd behaviour. The authors initialise their method by applying frame segmentation to extract a region of interest (ROI). Moreover, to extract flow fields, optical flow pruning is applied based on a predetermined threshold of a Euclidean length of their vectors. A combination of Moving Difference Image (MDI), Gaussian Mixture Model (GMM), K-means clustering, and Adaptive Median is used to achieve this. Figure 11, is a PETS2009 (Ferryman and Shahrokni, 2009) sample frame that is used as an input in the framework of this method. Lastly, the authors use a gradient ascent on the discriminative models previously noted with window sizes varying from 0-8 to model the flow-blocks pattern sequence. An anomaly is detected from the statistical ratio of anomalous flow-blocks and total flow-blocks. Experimentation of the method is applied on the PETS2009 (Ferryman and Shahrokni, 2009) dataset, and results of the recognition ratio are 96.2%, 97.1%, 98.1% for CRFs, HCRFs, and LDCRFs (on window size 3).

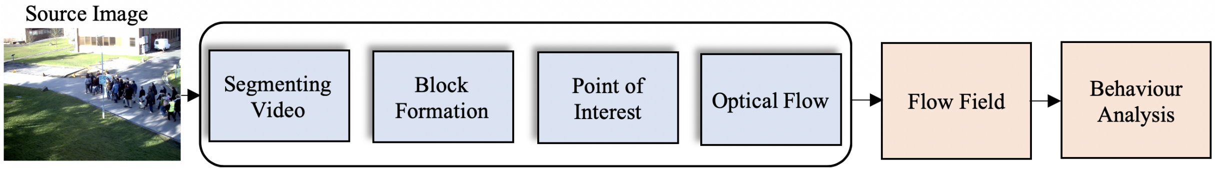


Figure 11: Crowd Behavior Analysis Using Discriminative Models Framework. Adapted from(El-Etriby et al., 2017)

Zhang et al. (2018a) propose an approach to detect anomalous behaviour within a crowd by extracting feature points, constructing a motion field, and applying an anomaly decider. By integrating the merits of SIFT features (Lowe, 2004) and Harris corner point, the authors have devised a multi-scale method to extract feature points. Using the successfully tracked feature points produced by the Lucas-Kanade algorithm, the crowd's motion field is constructed. The movement of well-tracked feature points is determined using speed and direction, quantified by the difference in spatial positioning between neighbouring frames. Using a predetermined threshold, a comparison to the motion field statistical information distribution is made to determine abnormality. Three scenes from the UMN dataset (University of Minnesota, 2006) were used to assess the algorithm's accuracy level. The abnormality threshold for both motion speed and motion direction was 40% and 0.7, respectively, and the decision threshold was 50%. The resulting anomaly detection rates were compared to results from social force model detection (Chen and Huang, 2013) and spatial-temporal motion statistical model (Li et al., 2014). The average anomaly detection rate from the three scenes was 98.33%, the spatial-temporal motion statistical model average detection rate was 97.1% and the lowest, social force model, averaged only 96.2%.

Real-time detection of anomalous actions within a low-medium density crowd is shown in (Bera and Manocha, 2018). To create a state representation of the crowd, the authors initially extract pedestrians' motion trajectories in the crowd using the Reciprocal Velocity Obstacles (RVO) approach (van den Berg et al., 2011). Current position, average velocity, cluster flow, and intermediate goal position of pedestrian or cluster of pedestrians are the trajectory-level features computed to analyse the crowd's behaviour. Bayesian inference algorithm is applied to estimate pedestrian states, resulting in an overall crowd state. An anomaly is detected when the Euclidean distance between the pedestrian's local features and their global features increases above a predetermined value. Area Under Curve (AUC) and Accuracy results are documented for testing on 879-44 (Rodriguez et al., 2011b) and ARENA (Patino and Ferryman, 2014) and UCSD (Chan et al., 2008) datasets: 0.97, 80%, 0.91, 76% and 0.873, 85% respectively.

3.3.3 Anomaly detection using Neural Networks

Mahadevan et al. (2010) utilise local video feature extraction to detect anomalies within a crowd. Instead of using global feature extraction such as Markov Random Field (MRF) or Latent Dirichlet Allocation (LDA), this research uses three local properties for video representation. The first is dynamic and appearance of crowd patterns using mixtures of dynamic textures (DTs), the second is temporal abnormalities extracted using Gaussian Mixture Model (GMM). The last is spatial abnormalities extracted using a saliency detection method. These representations are used to model a crowd's normal behaviour, detected outliers under this model are considered abnormalities. The research has shown that dynamic textures are more fitting than optical flow in the process of crowd anomaly detection. However, this method is computationally heavy; a frame of 240x160 requires 25 seconds of computation to test. Experimentation was applied on the UCSD dataset, and results are noted in Table 5.

Hasan et al. (2016) present a fully connected convolutional feed-forward deep auto-encoder network that learns regular motion patterns (regularity) from input videos. The idea of the method is; after training the network to reconstruct regularity, the network will not be able to reproduce irregular motion patterns accurately. Initially, the framework utilises Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) and Histograms of Optical Flows (HOF) (Dalal et al., 2006) to extract improved trajectory motion features. The regular motion patterns are used to learn an auto-encoder based on an end-to-end neural network. Then a fully convolutional auto-encoder is used to learn local features and the classifiers. This method's drawback is that reconstruction tends to give high anomaly detection scores for new normal patterns. An illustration of the given framework is shown in Figure 12. Experiments are applied for multiple applications such as learning temporal regularity, future frame prediction, and abnormal behaviour detection. Results for abnormal behaviour detection on Avenue and UCSD datasets are noted in Tables 6 and 5.

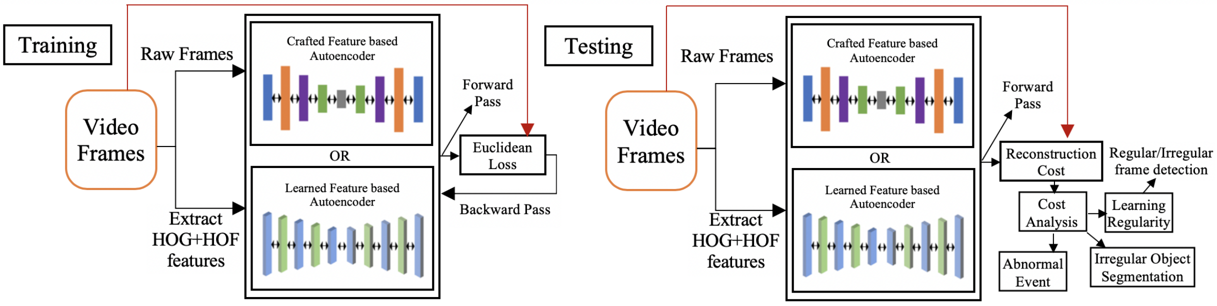


Figure 12: Learning Temporal Regularity in Video Sequences framework. Adapted from (Hasan et al., 2016)

Ravanbakhsh et al. (2016) capture abnormality in frame sequences by tracking alterations of CNN features throughout time. More accurately, a Fully Convolutional Network (FCN) is given frame sequences. A binary quantisation layer is then used to restrict the quantity of the high-dimensional feature maps (quantise) into compressed binary patterns. This binary quantisation layer is attached to the top of the FCN to generate binary maps. Spatial relationships of the input frame are protected throughout this process. Subsequently, a histogram is generated from a spatio-temporal block of the accumulated binary patterns. Finally, the output Temporal CNN Pattern (TCP) of the histograms merged with the extracted optical flow is used to detect abnormal regions. The complete framework is illustrated in Figure 13. Experiments were applied to UMN and UCSD datasets and noted in Tables 7 and 5. However, this network does not apply end-to-end training and requires a complex post-processing step.

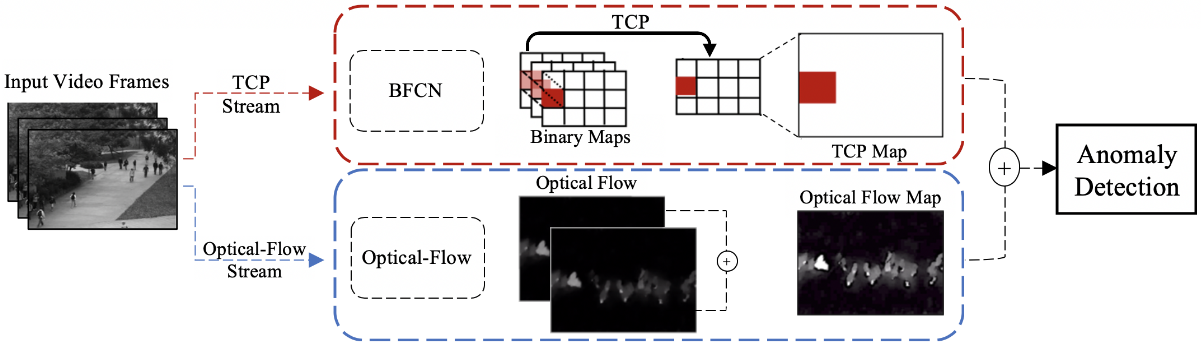


Figure 13: Plug-and-Play CNN for Crowd Motion Analysis framework. Adapted from (Ravanbakhsh et al., 2016)

Xu et al. (2017) present an Appearance and Motion Deep Network named AMDN. The network adopts multiple stacked denoising autoencoders (SDAEs) for feature representations. The feature representations are extracted by utilising a double fusion (early and late fusion) architecture that joins low-level motion and appearance features. More accurately, two SDAEs receive calculated optical flow and image patches as input to generate motion and appearance feature, respectively. Subsequently, an early fusion stage is used to merge frame pixels and the conforming optical flow to teach a third SDAE the joint representation of motion and appearance features. Several one-class support vector machine (SVM) models are trained on the extracted feature representations to calculate a set of anomaly scores. Finally, the late fusion stage is used to merge these anomaly scores to detect abnormalities. Experiments were applied to three datasets: UCSD, Subway and Train, results of USCD are noted in Table 5. The network is prone to over-fitting due to small abnormal training data and the small frame patches' restriction. Additionally, the network is not trained end-to-end, considered relatively shallow, and it requires that several multiple SVMs be trained externally.

Inspired by the work presented in Hasan et al. (2016) (documented above), the authors of Chong and Tay (2017) generate a video representation from a set of extracted general features. The method is semi-supervised and utilises a stack of convolutional autoencoders (AEs). The stack of convolutional AEs is used to process input video frames and extract spatial features. These features are then used as input into another stack of convolutional AEs for temporal feature extraction. As shown in Figure 14, the deep end-to-end network is trained on normal video frames. An anomaly is detected based on the reconstruction error between the input video frames and the reconstructed video frames. Low reconstruction error from the network indicates normal scenes, whereas a high reconstruction error indicates abnormal scenes. A threshold would determine the occurrences of abnormality within the footage. More specifically, the network has three main stages to detect anomalies; the first stage is a pre-processing stage to resize and scale the input video frames and divide the input frames into video volumes

(10 frames each). The second stage, feature learning, utilises a spatial encoder/decoder with a two-layer convolution and deconvolution to learn spatial features. Additionally, the temporal encoder is based on a three-layer convolutional long short term memory (LSTM) to learn temporal patterns of the spatial features. The third and final stage calculates the regularity scores based on all input video frames' reconstruction error to determine abnormality occurrences. The method results are documented in Tables 5 and 6, although the AUC and EER results surpass other methods, the network produces more false alarms than others.

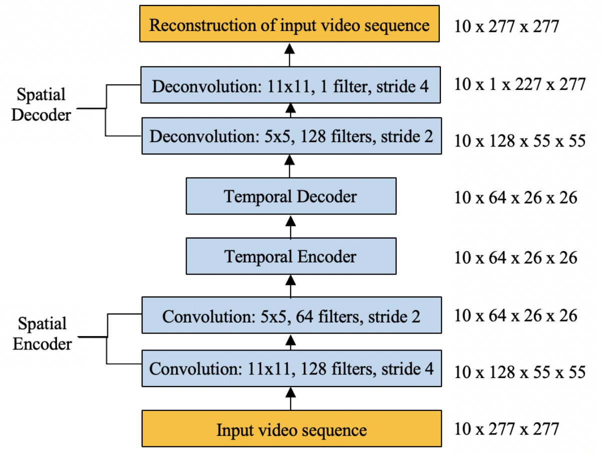


Figure 14: Stacked convolutional autoencoders with spatial and temporal encoder/decoder. Adapted from (Chong and Tay, 2017)

A *fully convolutional neural network* (FCN) is applied in Sabokrou et al. (2018), for the detection of anomalies in a crowd. Temporal data is extracted, and a pre-trained supervised FCN is fed to an unsupervised FCN to detect global anomalies within a crowd. A normal reference model is created using a fitted Gaussian distribution classifier, as shown in Figure 15. Additionally, to better represent abnormal regions, generated by the AlexNet (Krizhevsky et al., 2012), an auto-encoder is applied to the suspicious regions. Abnormal regions are established if they are not similar to the normal reference model. The FCN consists of two initial convolution layers using an adjusted version of AlexNet. The first layer is used to differentiate between normal and abnormal regions; this layer's output result contains many false positives. The second layer is a deeper discriminative layer, achieving better results. A final layer is used to attain better and deeper features, but the layer is likely to over-fit. Experimentation is applied on the UCSD (Chan et al., 2008) and Subway (Adam et al., 2008) benchmark datasets, the results are compared to other methods using the Area Under Curve (AUC), Receiver Operating Characteristics (ROC) curve and Equal Error Rate (EER). The proposed method outperforms the examined state-of-the-art methods with regards to quantitative measures and faster run-times. Details of the quantitative results are documented in Table 5.

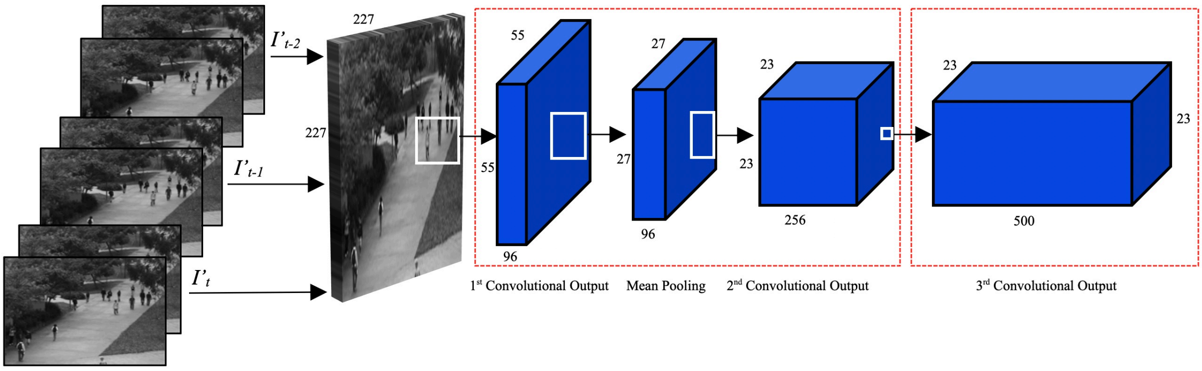


Figure 15: The FCN structure is applied for regional feature extraction. Two Gaussian classifiers are integrated in later stages to label abnormal regions. Adapted from (Sabokrou et al., 2018)

Majumder et al. (2018) use recurrent neural networks (RNN) to extract anomaly from motion. The framework uses two stacked LSTMs (Long short-term memory) as an encoder-decoder to define *normal* behaviour. The authors describe an anomaly as any motion that does not follow the normal pattern. Peculiarly, sudden movements and motions that are slower, faster, or in a different direction to the observed scene. Farnebäck (2003) algorithm is applied to extract the dense optical flow (further explained in Section 3.4.1) magnitudes for each scene. In the training process, three LSTM networks are trained on different scales. Sequences are formed from optical flow stacking and fed to the stacked RNNs to predict the future flow. Multiple datasets are used for testing, and the qualitative results are compared to other anomaly detection algorithms. Quantitative results are documented from testing on the UMN (University of Minnesota, 2006) dataset and produced an AUC value of 99%.

Similar to Majumder et al. (2018) the method applied by Qiu et al. (2018) use a Convolutional Neural Network followed by an LSTM (Graves et al., 2013) to detect anomalous objects. The framework combines the trajectory and motion-based techniques by extracting objects using CNN and feeding said data to the LSTM. The CNN applied is based on the you-only-look-once (YOLO) (Redmon and Farhadi, 2018) detector, which outputs a bounding box of the detected objects. Because object representation is simple, the method is computationally cheap and fast. The LSTM model applied considers not only spatial and/or temporal data of each object, but also includes correlation data about the objects' neighbours. Position, velocity, acceleration, and direction are all the characteristics used to interpret normalcy using a threshold. The CNN object extraction method is trained on ImageNet (Deng et al., 2009), and the LSTM uses the OTB-30 (Wu et al., 2013) dataset for training/testing. Comparative success plots indicate better performance than standard tracking techniques. Experimentation results are documented as one pass evaluation (OPE) of 0.467, temporal robustness evaluation (TRE) of 0.559 and spatial robustness evaluations (SRE) of 0.544.

Fan et al. (2020) use a partially supervised deep learning method to detect anomalies using normal samples. Dynamic flows, an amalgamation of multiple sequential optical flow frames, are produced using a Ranking SVM to consider long-term temporal data. The generated dynamic flows are fed into a two-stream Gaussian Mixture Fully Convolutional Variational Auto-encoder (GMFC-VAE). RGB images from the normal sample data are also fed into the GMFC-VAE. The GMFC-VAE utilises the feature representations: RGB images (appearance cues) and dynamic flows (motion cue) to detect anomalies. The encoder-decoder is based on a Fully Convolutional Network (FCN) which does not include a fully-connected layer. Respective spatial locations of the input image and output feature map are saved. Anomaly scores are given based on a sample energy method to test samples. The UCSD (Chan et al., 2008) and Avenue (Lu et al., 2013) (results noted in Table 6) datasets are used for experimentation, and evaluations are based on both frame-level and pixel-level criterion. Area Under Curve (AUC), Equal Error Rate (EER), True Positive Rate (TPR), and False Positive Rate (FPR) are all used to evaluate the system. Regarding frame-level detection, the proposed system outperforms systems such as (Mehran et al., 2009; Sabokrou et al., 2018), as shown in Table 5.

3.3.4 Crowd anomaly detection using Generative Adversarial Networks

GANs are typically used to generate fake data that can be construed as real data, the framework has been utilised in applications such as image classification, images generation, and image classification. Only recently has the framework been utilised for crowd anomaly detection.

Ravanbakhsh et al. (2017) have utilised the framework for that purpose, the generative network is used to model normal data. With the lack of availability of abnormal datasets, the generator has the benefit of being trained on only normal data. Abnormal data is then detected by measuring the distance between the generated and the learned data. More specifically, the authors used the framework presented by Isola et al. (2017) to learn the translation between optical flow (further explained in Section 3.4.1), computed using (Brox et al., 2004), and the corresponding input frames. The framework of the method is illustrated in Figure 16, when testing the network would not be able to generate abnormal scenes because it is trained on normal footage and local difference is used to detect anomalies. In the experimental setup, testing was applied to both UCSD (Table 5)and UMN datasets (Table 7). Quantitative results are documented with respect to frame-level and pixel-level abnormality detection/localisation. In comparison to methods proposed by Ravanbakhsh et al. (2016) and Xu et al. (2017), this method has shown better AUC and EER values. The disadvantage of this architecture is the dependency on a CNN, pre-trained on ImageNet, to collect an adequate amount of semantic data. The AUC and EER results from testing on both datasets are documented in Table 5.

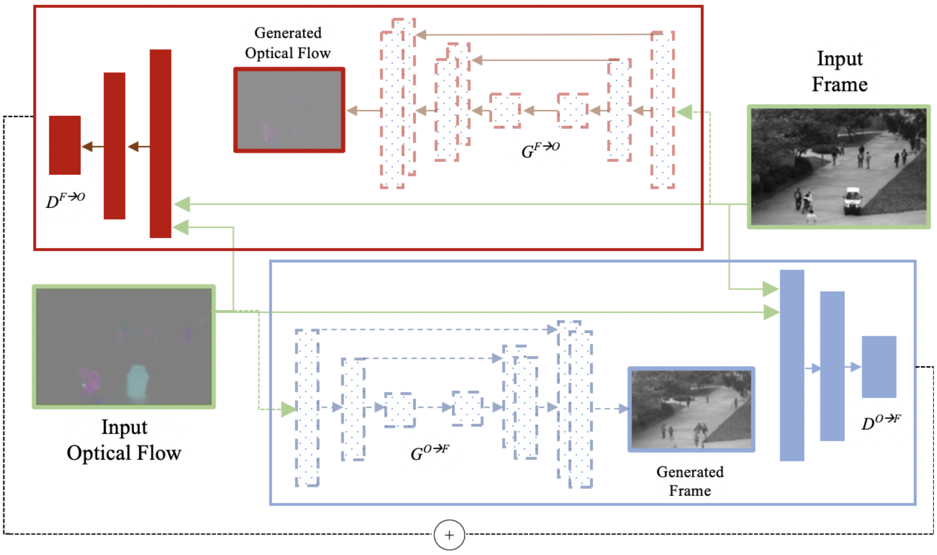


Figure 16: Adversarial discriminators for crowd abnormal event detection. Adapted from Ravanbakhsh et al. (2019)

Ravanbakhsh et al. (2019) present a continuation of the previously discussed work. Unlike Ravanbakhsh et al. (2017), the discriminative network is used to detect anomalies during testing, whereas their previous work used the generative networks' reconstruction errors for anomaly detection. The same testing setup is applied, using the same datasets. The AUC result produced when testing this architecture on the UMN dataset is 0.99 (Table 7). This result is very similar to the result of 0.99 from their previous work on the same dataset. While the authors' preceding work depends on a pre-trained CNN for semantic data and a fusion strategy to consider both pixel-level and semantic-based reconstructions errors, their latter work does not appear to be quicker with regards to training time. However, the authors claim that testing time is reduced due to the use of the adversarial discriminator for detection. The results of testing on UCSD datasets are presented in Table 5.

A novel approach to anomaly detection was presented by Liu et al. (2018b) where the authors use future frame prediction for the purpose of detecting anomalous behaviour. The fundamental idea of this approach is to utilise the deviation between ground truth video frames and their corresponding predicted future frames to find irregular scenes. The network, as shown in Figure 17, generates a prediction frame from a U-Net and utilise different constraints to achieve higher quality frame prediction. Some of the constraints used are adversarial training loss (Isola et al., 2017) (further described in Section 4.2) to better train the model in generating better quality images. The second constraint used is a spatial constraint based on intensity and gradient loss. Lastly, an optical flow loss calculated using a pre-trained network Flownet (Dosovitskiy et al., 2015) is used as the motion constraint. The complications of this method are that Flownet is considered costly for optical flow extraction and the network produces high false-positive rates. The network is trained and tested on multiple datasets: Avenue, UCSD and ShanghaiTech. The results of UCSD and Avenue are presented in Tables 5 and 6.

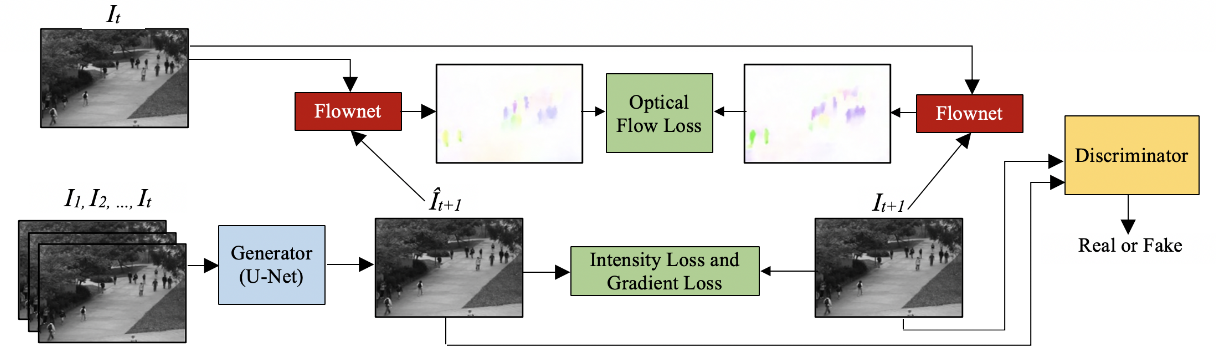


Figure 17: Future frame prediction for anomaly detection framework. Adapted from (Liu et al., 2018b)

The authors of Vu et al. (2019) present MLAD (MultiLevel Anomaly Detector), a network based on the anomaly detection system by Ravanbakhsh et al. (2017) with additional denoising autoencoders (DAEs). A DAE is a neural network trained to reconstruct data from input data which is intentionally corrupted (noise is introduced). The multilevel representation system utilises low-level and abstract-level features to detect anomalies. Initially, the network is trained by calculating the optical flow frames corresponding to the input frames. Two DEAs are separately trained on both the input video frames and the matching calculated optical flow. The trained DEAs then extract high-level features for the input and optical flow frames. The generated features (high-level and motion) are fed into a conditional generative adversarial network (CGAN) (Isola et al., 2017) to train the network. For testing, the optical flow is calculated, the high-level features are generated using the pre-trained DAEs and the trained CGANs generate various error maps. Lastly, binary detection maps are deduced from the error maps and collated to produce a detection outcome. Experiments are applied to the USCD and Avenue datasets and the results are presented in Tables 5 and 6.

More recent research conducted by Pourreza et al. (2021a) consider the irregularity detection problem as a binary classification method. With the use of Wasserstein GANs, the authors train the generative network in the typical manner where the generator is trained on normal samples (normal behaviour). However, while training the generator on normal data, any failed normal data produced by the generator is considered as abnormal data. The produced abnormal data and normal data are then used to train a binary classifier for the detection of irregular samples. The authors used this method for both video anomaly and image outlier detection. They test this framework on UCSD Ped-2 for frame-level video anomaly detection and produce an EER result of 11% (Table 5).

Yu et al. (2021) create a novel abnormal event detection model named Adversarial Event

Prediction (AEP). The model is used for the detection and localisation of anomalies within crowds based on event prediction. Similar to other abnormal detection models based on GANs, the AEP model is trained on normal samples. However, the AEP contains the generator and three discriminators: latent feature, future, and past discriminator. The latent feature discriminator is used to derive the distribution of latent features to normal distribution. Whereas the future and past discriminators are used to differentiate between future and past events. Moreover, unalike the typical optical flow methods used for temporal development extraction, the authors build the generator, future discriminator, and past discriminator using 3-D CNNs (Ji et al., 2012) and fully connected neural networks. These neural networks are used for spatial and temporal extraction as well as abstracting the learned extractions. However, these 3-D CNNs require high computational cost. The AUC and EER results from testing on both the UCSD and Avenue datasets are documented in Tables 5 and 6.

Table 5: Experimental results of the state-of-the-art on UCSD, ERR (Equal Error Rate) and AUC (Area Under Curve) for frame and pixel level detection are documented.

	Dataset							
	Ped-1				Ped-2			
Method	Frame Level		Pixel Level		Frame Level		Pixel Level	
	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)
1. Social Force	67.5	31	19.7	79	55.6	42	-	80
2. MDT	81.1	25	44.2	58	82.9	25	-	55
3. Detection at 150fps	91.8	15	63.8	43	-	-	-	-
4. Plug-and-Play	95.7	8	64.5	40.8	88.4	18	-	-
5. ConvAE	-	-	81	27.9	-	-	90	21.7
6. AMDN	92.1	16	67.2	40.1	90.8	17	90.8	17
7. GAN generative	97.4	8	70.3	35	93.5	14	-	-
8. ConvLSTM	89.9	12.5	-	-	87.4	12	-	-
9. AnoPred	83.1	-	-	-	95.4	-	-	-
10. Deep-anomaly	-	-	-	-	-	11	-	15
11. MLAD	82.34	23.5	66.6	**22.65**	**97.52**	**4.68**	**94.45**	**4.58**
12. GAN discriminative	96.8	7	70.8	34	95.5	11	-	-
13. Gaussian Mixture	94.9	11.3	71.4	36.3	92.2	12.6	78.2	19.2
14. G2D	-	-	-	-	-	11	-	-
15. AEP	**97.92**	**6.07**	**74.83**	31.06	97.31	7.52	-	-

The results presented in the comparison table above are from the methods by 1. (Mehran et al., 2009), 2. (Mahadevan et al., 2010), 3. (Lu et al., 2013), 4. (Ravanbakhsh et al., 2016), 5. (Hasan et al., 2016), 6. (Xu et al., 2017), 7. (Ravanbakhsh et al., 2017), 8. (Chong and Tay, 2017), 9. (Liu et al., 2018b), 10. (Sabokrou et al., 2018), 11. (Vu et al., 2019), 12. (Ravanbakhsh et al., 2019), 13. (Fan et al., 2020), 14. (Pourreza et al., 2021a), 15. (Yu et al., 2021). The best results are indicated in bold lettering.

Table 6: Experimental results of the state-of-the-art on Avenue dataset, ERR and AUC for frame level detection are documented.

Method	Frame Level	
	AUC (↑)	EER (↓)
3. Detection at 150fps	80.5	-
5. ConvAE	70.2	25.1
8. ConvLSTM	80.3	20.7
9. AnoPred	84.9	-
11. MLAD	71.54	36.38
13. Gaussian Mixture	83.4	22.7
15. AEP	**90.2**	**10.07**

Table 7: Experimental results of the state-of-the-art on UMN dataset, AUC results are documented.

Method	AUC (↑)
1. Social Force	0.96
4. Plug-and-Play	0.988
7. GAN generative	**0.99**
12. GAN discriminative	**0.99**

3.3.5 Summary

The previously discussed work on behaviour analysis is continually improving. Significantly the most recent work shows promise to what can be achieved within this discipline. Further improvements are still required as the experimental results are not satisfactory enough to be applied to the real-world environment. Additionally, the limitations of the crowd density are very clear in the discussed work; high-density crowds are not targeted as much due to its difficulty in application. Amongst the work presented, there is also a noticeable gap regarding the use of multiple views for behaviour analysis. Both the handcrafted approaches as well as the neural network approaches suffer from a lack of applicable "abnormal" behaviour datasets to train/test. Generative Adversarial Networks offer a promising solution as they can be trained on just "normal" behaviour datasets and as shown in Tables 5, 6 and 7. GANs (Ravanbakhsh et al. (2017), Ravanbakhsh et al. (2019), Liu et al. (2018b), Vu et al. (2019), (Pourreza et al., 2021a), and (Yu et al., 2021)) have proven to outperform other state-of-the-art methods. Detection and localisation results (EER and AUC) in Vu et al. (2019) and Yu et al. (2021) have demonstrated leading performance. Furthermore, most methods tend to utilise high accuracy optical flow estimation (Brox et al., 2004) for temporal feature extraction. Contemporary methods for optical flow estimation (Sun et al., 2017) or dynamic image extraction (Bilen et al., 2016) (detailed in Section 3.4) should be utilised for temporal feature extraction. The aforementioned temporal feature extraction methods have excelled in the field of action recognition. Section 3.4

investigates the different action recognition methods (temporal feature extraction) that can be integrated with CGANs as a novel approach for crowd anomaly detection.

3.4 Action Recognition

Action recognition is a method that extracts video frame features to enable the classification of actions according to class labels. Human action recognition is an important topic within fields such as robotics, surveillance and human-computer interaction. The standard framework flow for action recognition usually includes feature extraction, action learning, action segmentation, action classification and action model database (Herath et al., 2017). Moreover, two main representations of action recognition utilise holistic and/or local representations. For the purposes of this research, the main focus is feature extraction for action recognition. The feature extraction block extracts Holistic and Local representation and can be further used for crowd anomaly detection. Optical flow estimation methods (Brox et al., 2004; Sun et al., 2017) and dynamic image extraction methods (Bilen et al., 2016, 2017) are investigated below.

3.4.1 Optical Flow

Optical flow is defined as the estimation of the temporal (motion) development of every pixel from several consecutive input frames. The extracted motion patterns are based on the movement of objects, surfaces and edges. The estimated flow for two consecutive frames is usually represented by a vector field where each pixel of the first frame is associated with a displacement vector to determine its location in the second frame (Horn and Schunck, 1981). Conducted experimentation and graphical representations of optical flow estimation methods on two consecutive frames are shown in Section 6.2. Both a standard optical flow method (Brox et al., 2004) and a more novel state-of-the-art method for optical flow estimation (Sun et al., 2017) are presented hereafter.

3.4.1.1 Standard Optical Flow

Brox et al. (2004) utilise an energy-based optical flow estimation method built around three constraints. The method assumes a brightness consistency, a gradient consistency and a discontinuity-preserving displacement smoothness assumption. Input frames are given to the model and the method assumes that the grey value of each pixel does not change after it has been displaced. However, due to brightness variability from one frame to the next the grey value is susceptible to minimal disparities. The gradient consistency constraint is introduced because it does not change if the grey value changes. The gradient constraint also assumes that the pixel gradient does not change after displacement. Lastly, due to the previous constraints being applicable in a local fashion, without consideration to the relationships of the neighbouring pixels, the smoothness constraint is introduced. The constraint can overcome some of the outlier estimates by assuring continuity in the flow field. By considering all three constraints, the method can estimate the optical flow field between consecutive frames more accurately.

3.4.1.2 Novel Optical Flow

The optical flow model presented by Sun et al. (2017) has established performance results (increased accuracy, reduced model size, reduced training and running time) that outperform other optical flow algorithms. The authors use a CNN model that is based on pyramidal processing, warping, and the use of cost volume. The network, shown in Figure 18, begins by extracting raw images and casting learnable feature pyramids as an alternative to the fixed image pyramid, this is done because consecutive images can be different due to light and shadow modifications. The second task is to apply a conventional warping method for significant motion estimation as a layer in the network. The third step is a network layer that builds a cost volume to be processed and utilised in flow estimation. Cost volume is the processes of storing data matching costs of pixels and their equivalent pixels in the following image frame. Lastly, contextual data is extracted and used to further enhance the produced optical flow. The conducted practical application of this method is documented in Section 6.2

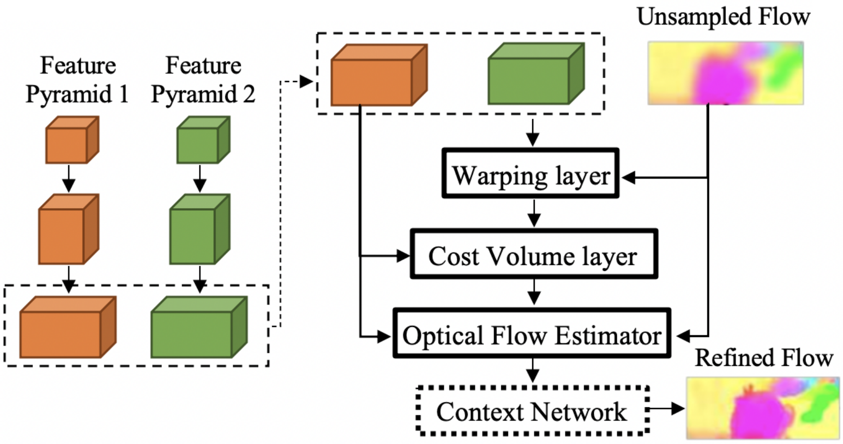


Figure 18: Overview of the PWC-Net optical flow framework. Adapted from (Sun et al., 2017)

3.4.2 Dynamic Images

Dynamic Images is a method proposed by Bilen et al. (2016) that represents a group of consecutive frames (videos) as a single RGB image. With the use of CNNs, the algorithm uses rank pooling and the dynamic image is produced through the ranking machine's parameters. The ranking machine encodes the temporal development data extracted from the image frames. While computing the dynamic image the applied ranking classifier replaces the usage of feature representation data and instead uses frame pixels. To increase efficiency, simple linear operations are applied to the images in the rank pooling process. The result of this algorithm is a single RGB image that represents the entirety of the inputted frames (input video). The conducted practical application of this method is documented in Section 6.2.3 and an example shown in Figure 19.

A continuation of Bilen et al. (2016)'s work is presented in Bilen et al. (2017) where the network architecture is extended to use two additional streams yeilding a four-steam framework for action recognition. A representation of the four-stream architecture is illustrated in Figure 20. The streams include a single image, dynamic image, optical flow and dynamic optical flow to predict the action presented in video frames. Another extension includes fine-tuning the network to increase accuracy with regards to recognising actions. Table 8 notes the accuracy results of optical flow and dynamic images for the purposes of action recognition on the UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011) datasets.

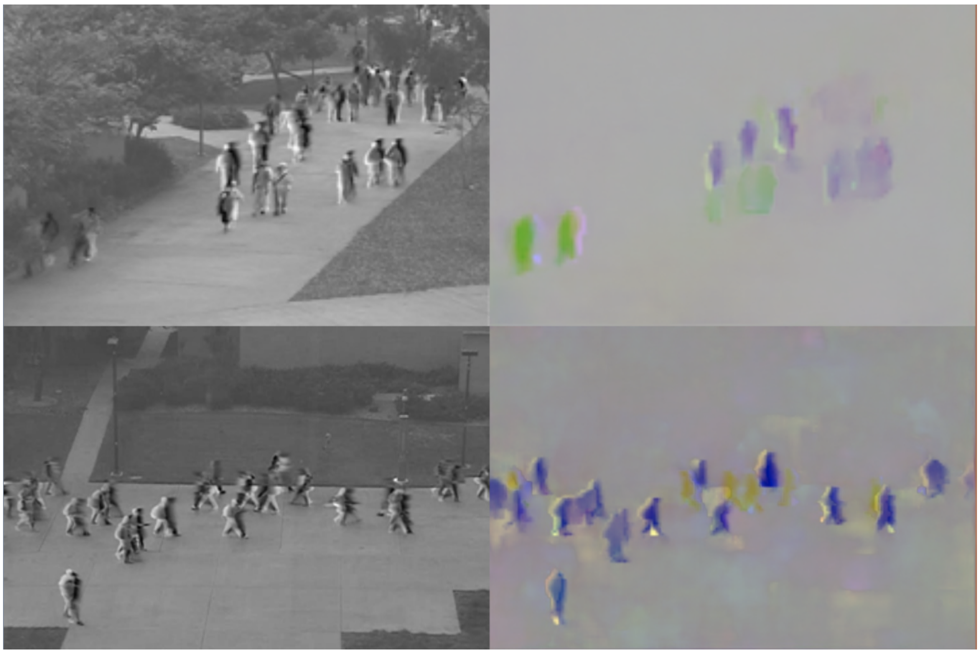


Figure 19: Sample Dynamic Image (left) and Dynamic Optical Flow Image (right) results on UCSD dataset.

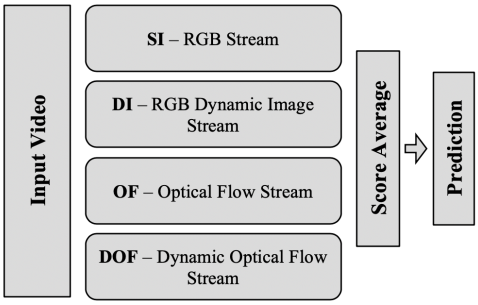


Figure 20: Four-stream architecture for action recognition. Adapted from (Bilen et al., 2017)

There are various action recognition methods that utilise temporal features such as histograms of optical flow (HOF) in addition to other features to estimate an action. For the purposes of this research, focus is given to methods that use only optical flow for action recognition. As shown in Table 8, the method presented by Simonyan and Zisserman (2014) use a temporal stream (optical flow features) when evaluating their method, it is evaluated in comparison to Dynamic Images (Bilen et al., 2016, 2017). Dynamic image for action recognition have proven to achieve higher mean class accuracy (the evaluation metric used in action recognition to determine accuracy level) than optical flow.

Table 8: Mean Class Accuracy results on UCF101 and HMDB51 datasets for action recognition.

	Dataset	
Method	HMDB-51	UCF-101
Temporal stream ConvNet (Simonyan and Zisserman, 2014)	54.60	83.70
Dynamic Image (Bilen et al., 2016)	57.3	**86.6**
Dynamic Optical Flow (Bilen et al., 2017)	**58.9**	**86.6**

The sections below provide more details of the evaluation metrics and datasets referred to in the above discussions of previous work. These will also be used in subsequent chapters.

3.5 Evaluation metrics

Evaluation metrics can either be qualitative or quantitative; qualitative is merely based on an examination of visual results while quantitative metrics are tangible measurements. The following metrics are used to distinguish the performance between the different algorithms used in crowd analysis. It is vital for researchers to be able to evaluate techniques in a quantitative manner. Some of the most commonly used performance quantitative metrics used are ROC curves, Accuracy, Recall, Precision, and Error rates. Metrics used for evaluation are presented below:

- **Accuracy** is used to evaluate the correctness of an algorithm and is calculated using the equation below:

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

True Positives (TP), True Negatives (TN), Total of positives (P), and Total of Negatives (N).

- **Recall** metric r equates to the ratio of the number of positive samples that are appropriately classified and the total quantity of samples that are truly positive (true positives and false negatives). Recall is stated as:

$$Recall(r) = \frac{TP}{TP + FN} \tag{2}$$

True Positives (TP), and False Negatives (FN).

- **Precision** metric p equates to the ratio of the number of positive samples that are appropriately classified and the total quantity of positives samples. Precision is stated as:

$$Precision(p) = \frac{TP}{TP + FP} \tag{3}$$

True Positives (TP), and False Positives (FP).

- **F1 Score** is a combination of both the Recall (r) and the Precision (p) and is denoted as

$$F1 = \frac{2pr}{p + r} \tag{4}$$

Precision (p), and Recall (r).

- **Mean Square Error and Mean Absolute Error:** are both evaluation metrics used to assess the quality of estimators; for the purposes of this research, they are used to evaluate crowd counting techniques.
  - **Mean Square Error**

$$\epsilon_{sqr} = \frac{1}{N} \sum_{N}^{i=1} (y_i - \hat{y}_i)^2 \tag{5}$$

Number of test frames (N), the ground truth number (y_i), and the number projected from the ith frame ($\hat{y}_i$).

  - **Mean Absolute Error**

$$\epsilon_{abs} = \frac{1}{N} \sum_{N}^{i=1} |y_i - \hat{y}_i| \tag{6}$$

Number of test frames (N), the ground truth number (y_i), and the number projected from the ith frame ($\hat{y}_i$).

- **ROC Curves**, Receiver Operating Characteristic curve, is illustrated as a graphical plot of the values of **True Positive Rate (TPR)** alongside the values of **False Positive Rate (FPR)** at variable thresholds, the equations being denoted as:

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

True Positive Rate (TPR), False Positive Rate (FPR), True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN).

- **Equal Error Rate**
  The Equal Error Rate (EER) is the point on the ROC curve where False Positive Rate (FPR) is equal to (1 − True Positive Rate (TPR)), or where their curves intersect. The lower the EER value is the higher the performance.

- **Area Under the ROC Curve**
  Area under the ROC Curve (AUC) measures the entire two-dimensional area under the entire ROC curve. As shown in Figure 21 AUC is the area under the ROC curve and the EER is the specific point in the ROC curve.

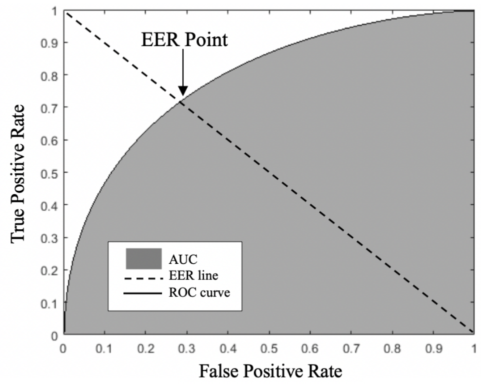


Figure 21: Illustration of AUC, ROC and EER evaluation metrics.

These basic evaluation metrics are utilised through the various fields of crowd analysis (crowd counting, crowd tracking, motion representation and crowd anomaly detection). The quantitative evaluation metrics should be used throughout experimentation to allow consistent performance assessment between the variety of algorithms presented by researchers. More specifically, for abnormal behaviour detection within a crowd ROC, EER and AUC are consistently utilised in experimentation.

3.6 Datasets

Datasets are very important to the analysis of crowd behaviour. More specifically, benchmark datasets are a necessity; researchers can apply, compare and evaluate their framework against others when the datasets are consistent in experimentation. For the purposes of this research, datasets related to behaviour understanding, crowd counting, crowd recognition, crowd segmentation, crowd tracking, event detection, behaviour understanding, and abnormal behaviour detection are explored.

3.6.1 Crowd Anomaly Datasets

The definition of an "Anomaly" within a crowd is difficult to explain in a definitive way. The typically used general definition of a crowd anomaly is similar to the outlier detection problem (Mahadevan et al., 2010). Basically, an anomaly is any event that does not conform to the defined normalcy, when an anomaly occurs the corresponding video frames will be significantly different in relation to the older video frames (Chong and Tay, 2017). The defined normalcy is usually based on a learnt model established from "Normal" videos in anomaly detection datasets. Examples of "Abnormal" behaviour in benchmark datasets include behaviours such as:

- Throwing objects (papers or bags), unusual direction movement, presence of unusual objects (bikes or bags) (Avenue dataset (Lu et al., 2013)).
- Pedestrians moving in the opposite direction of the majority of people, loitering and irregular interactions (Subway dataset (Adam et al., 2008)).
- Presence of objects such as bikers, wheelchairs and small carts within a usually pedestrian-filled environment (UCSD dataset (Chan et al., 2008)).
- Quick dispersion of people in different directions (UMN dataset (University of Minnesota, 2006)).

3.6.2 Benchmark Datasets

Table 9 presents the details of all relevant datasets reviewed in this research. The datasets used frequently by the majority of crowd analysis methods in their experimentation and comparisons are discussed below. The datasets used specifically for crowd anomaly detection are highlighted in bold.

- **Avenue (Lu et al., 2013): the Avenue dataset contains 16 training videos and 21 testing videos captured in CUHK campus. Some of the abnormalities defined in this dataset are Strange action, Wrong direction and Abnormal objects. The resolution of the footage is 640x360 and ground-truth data is saved in Matlab format.**
- CUHK dataset (Chinese University of Hong Kong) (Shao et al., 2014, 2017): the CUHK dataset comprises of 474 videos of indoor and outdoor scenes, the variety of footage it what made this dataset popular. The resolutions of the footage vary from 240x352 to

1080x1,920 and the videos show different illuminations, occlusions, numerous densities and perspective scales.

- **BEHAVE (Blunsden and Fisher, 2010): contains two views of 10 acted-out scenes of low-density crowds. The acted-out scenes are categorised into Group, Approach, Walk Together, Split, Follow, Chase, Fight, Run Together, and Meet.**

- PETS 2009 (Ferryman and Shahrokni, 2009): another acted-out dataset is the PETS 2009, this is one of the rare datasets that capture footage from multiple angles while noting the camera calibration data. Additionally, the footage is divided for multiple purposes; training data, count and density estimation, tracking, and event recognition.

- **Subway (Adam et al., 2008): contains two captured surveillance events: exit gate and entrance gate. The captured videos are grey scale with a resolution of 512x384. The videos contain 209,150 frames and some of the anomalous behaviours include moving in the wrong direction, no payment, loitering and irregular interactions.**

- UCF datasets (Idrees et al., 2013; Ali and Shah, 2007, 2008): UCF has many public datasets; three of them are of use to the analysis of crowd behaviour. All three datasets are high in crowd density and are collected from online sources. The crowd counting dataset is only 50 images but has 64k manual annotations, the segmentation dataset is 38 videos, and the tracking dataset consists of 1289 images.

- **UMN dataset (University of Minnesota, 2006): eleven videos are captured from three different locations. The footage starts with a medium-density crowd of people acting out "normal" movement. After some time the crowd members suddenly disperse in different directions as if panicked.**

- **UCSD (Chan et al., 2008): An elevated stationary camera was used to collect footage of a medium-density pedestrian walkway. The dataset has 50 training and 48 testing videos in total. Ground truth annotation is included in every frame and binary value is used to indicate if an anomaly is present.**

- Violent-flows (Hassner et al., 2012): includes footage of both crowd violence and non-violence; the real-world videos were extracted from YouTube. This dataset is a benchmark to test violent/ non-violent crowd behaviour classification and recognition of violent occurrences.

- WWW Crowd Attribute dataset (Shao et al., 2015): this is one of the largest public datasets; it consists of 10,000 videos from 8,257 environments. The footage is annotated with 94 different attributes.

Sample images from the aforementioned datasets are displayed in Figure 22.

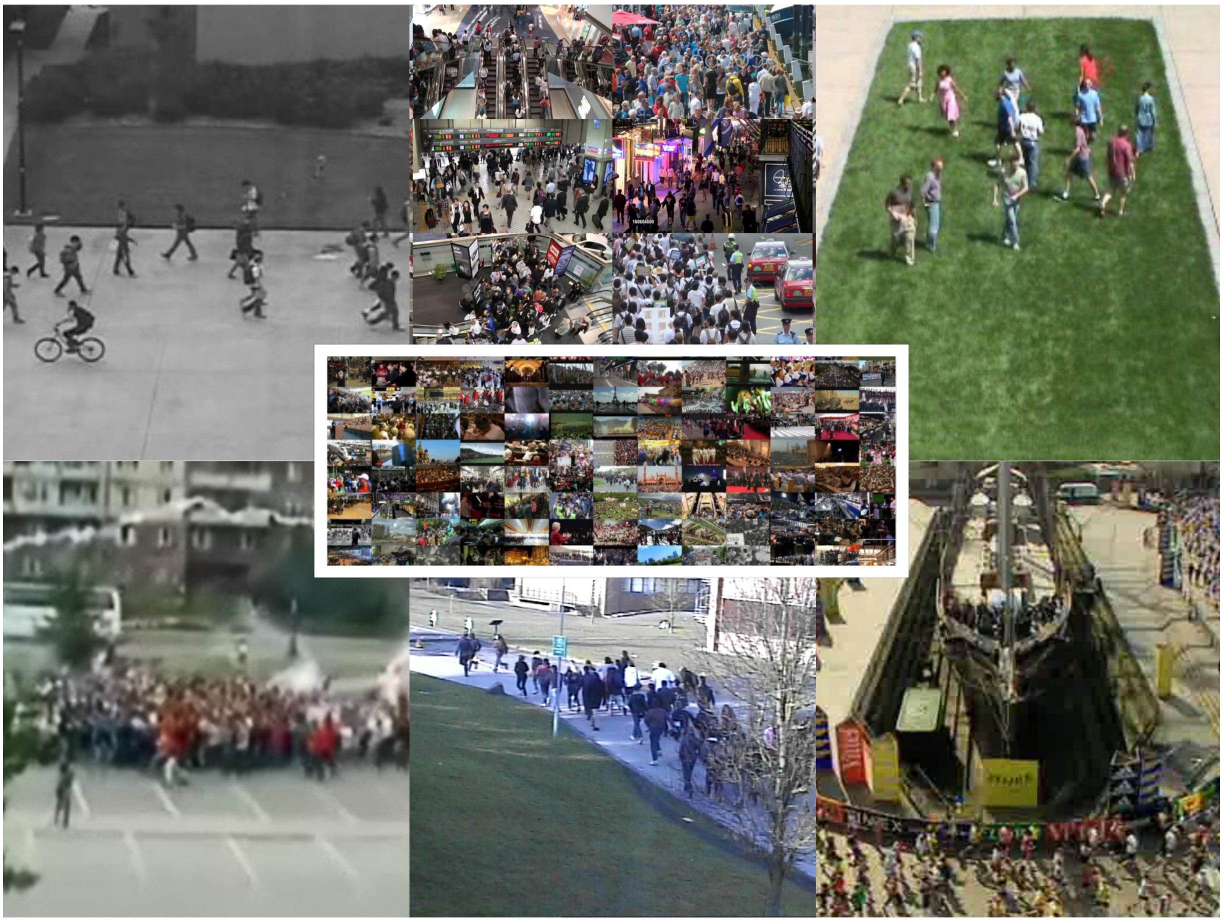


Figure 22: Sample images from the most prominent crowd datasets (from top left to bottom right): UCSD, CUHK, UMN, Violent-Flows, PETS 2009, UCF Dense-Tracking, and WWW Crowd Attribute (middle).

3.6.2.1 Summary

The survey of datasets highlights a substantial gap regarding datasets with combined features of high-density crowds, annotations and occurrences of anomalous behaviour. To overcome this lack of availability, a new annotated, high-density crowd dataset has been created and contains both normal and abnormal footage (anomalous behaviour). Collection and labelling of footage for this dataset is described in Chapter 5; the collected data adheres to the aforementioned features.

Table 9: Notable benchmark datasets.

Dataset	**Res. & Colour**	**Purpose**	**Description**
AVSS AB (Advanced Video and Signal based Surveillance, 2007)	720x576, RGB	Event detection	This dataset includes footage from different locations within the UK. Two scenarios are of concern, "Abandoned Baggage" and "Parked Vehicle".
BEHAVE (Blunsden and Fisher, 2010)	640x480, RGB	Behaviour Understanding	The BEHAVE dataset contains two views of 10 acted-out scenes. They are categorised as In Group, Approach, Walk Together, Split, Follow, Chase, Fight, Run Together, and Meet.
CAVIAR (CAVIAR 2003)	384x288, RGB	Abnormal Behaviour Detection	The CAVIAR dataset comprises of two events; the first was filmed at the entrance of INRIA Labs at Grenoble, France with the use of a wide-angle lens. The other was filmed from two different views: along and across a hallway in a shopping centre located in Lisbon, this was also shot using a wide-angle camera lens.
Chinese University of Hong Kong (CUHK) (Shao et al., 2014, 2017)	Varies, RGB	Holistic crowd motion	This dataset contains 474 videos from 215 crowded scenes. The footage has variable resolutions starting from as low as 240x352 to as high as 1,080x1,920. The different illuminations, occlusions, numerous densities and perspective scales, and the fact that there is footage of indoor and outdoor scenes makes this dataset unique.
Crowd Dataset + Ground Truth (Lim et al., 2014)	Varies	Crowd saliency	This dataset is one of the rare datasets that include a high-density level of crowds. The images vary in scenes and display behaviours that are both clear and that contain minor variability. No anomalous behaviours are included.
Crowd-11 (Dupont et al., 2017)	Varies, RGB	Behaviour Understanding	Dataset is created including definitions of daily life behaviours. Crowds are categorised into two types: Dynamic and No perceivable stream. Furthermore, there are ten classes to define the data. Resolutions vary from 220x400 to 700x1250.
Data-Driven Crowd Analysis dataset (Rodriguez et al., 2011b)	Varies, RGB	Holistic crowd motion	Footage from web pages such as Getty Images and YouTube was collected to create the Data-Driven Crowd Analysis dataset. The search engines were queried with words like "cross-walk", "festival", and "marathon". The resolutions vary from 480x360 to 720x480 additionally; there is footage of locations in the night time.
Grand Central Station dataset (Zhou et al., 2012)	720x480, RGB	Group motion recognition	The footage from this dataset is recorded at the Grand Central Station in New York, USA. The key point KLT trajectories obtained from the footage are annotated.

Table 9: Notable benchmark datasets.

Dataset	Res. & Colour	Purpose	Description
Mall dataset (Chen et al., 2012)	640x480, RGB	Crowd Counting and Profiling	The Mall dataset was obtained from a webcam installed in a shopping centre. The footage is publicly accessible and presents various challenges such as variable density levels, severe occlusions, changing movement levels, and perspective distortions. The datasets have over than 60,000 labelled pedestrians within the 2,000 video frames.
PETS 2009 (Ferryman and Shahrokni, 2009)	Varies, RGB	Behaviour Understanding	PETS2009 contains footage of various activities, filmed with the use of multiple cameras, and acted out by nearly 40 actors. It includes camera calibration data and is divided into four partitions: training data, footage for person count/density estimation, footage for people tracking, and footage for flow analysis/event recognition. The resolution of the footage is either 720x576 or 768x576.
PETS 2014: ARENA dataset (Patino and Ferryman, 2014)	Varies, RGB	Behaviour Understanding	The ARENA dataset has 22 acted out scenes, aimed at understanding the behaviour of a human around a stationed vehicle, it contains footage from a crossing path/ car park at the University of Reading. The targeted understandings are any action with the possibility of being a threat. The threats are categorised into three divisions: 'Something is wrong', 'Potentially criminal', and 'Criminal behaviour'. Four cameras capture footage of the environment around the vehicle with a resolution of 768x576. Another four cameras capture footage within the vehicle with a high resolution of 1280x960.
ShanghaiTech dataset (Zhang et al., 2016b)	Varies, RGB	Crowd Counting	ShanghaiTech dataset is a large-scale crowd counting dataset containing 1,200 images with 330,00 manually labelled heads. The images are varied in regard to viewpoints and situation occurrences within the scene. The dataset is divided into two parts: Part A and Part B; collected from the internet and footage of busy streets in Shanghai.
Subway dataset (Adam et al., 2008)	512x384, Grey Scale	Abnormal Behaviour Detection	The dataset contains captures surveillance footage from two locations: exit gate and entrance gate of a subway. The videos are two hours long and contain 209,150 frames, some of the anomalous behaviour is moving in the wrong direction, no payment, loitering and irregular interactions. The normal footage is the first 15 minutes of the footage.
The Unusual Crowd Activity dataset (UMN) (University of Minnesota, 2006)	320x240, Varies	Abnormal Behaviour Detection	The Unusual Crowd Activity dataset includes captured footage from three different locations. The footage begins with a number of people acting out normal movement in that environment then suddenly sparse in different directions as if panicked.

Table 9: Notable benchmark datasets.

Dataset	Res. & Colour	Purpose	Description
UCF Crowd Counting dataset (Idrees et al. 2013)	Varies, Grey Scale	Crowd Counting	The UCF Crowd Counting dataset is mainly used for crowd counting purposes; this is due to the density of the crowds within the images. The authors collected the images using FLICKR as their main source; they used words such as 'concerts', 'stadiums', 'protests', 'pilgrimages', and 'marathons' to find appropriate images. The dataset consists of 50 images with around 64K human annotations, varying in both resolution and number of individuals for each image. The number of people in an image is 94 to 4543 people with an average of 1280 individuals per image.
UCF Crowd Segmentation dataset (Ali and Shah, 2007)	Varies, RGB	Crowd Segmentation	In the UCF Crowd Segmentation dataset, there are 38 videos of human crowds as well as high density moving objects. There is a variety in the resolution, the number of people, the perspective, as well as the illumination in each video. The authors collected the data from two sources; the BBC Motion Gallery and Getty Images.
UCF Crowd Tracking dataset (Ali and Shah 2008)	Varies, RGB	Crowd Tracking	The UCF Crowd Tracking dataset is a collection of photos from FLICKR; the images are of three high crowd dense marathons. The total number of images is 1289 and the image resolutions vary between 480x360 and 720x404.
UCSD (Chan et al. 2008)	238x158, Grey Scale	Abnormal Behaviour Detection	An elevated stationary camera was used to collect the footage in the UCSD anomaly detection dataset; the environment was a walkway for pedestrians. The dataset is divided into Ped-1 and Ped-2 each being from a different perspective. Ped-1 is further subdivided into 34 training videos and 36 testing videos meanwhile Ped-2 contains 16 training videos and 12 testing videos. Ground truth annotation is included in every frame for all clips; it is denoted as a binary value to indicate the presence of an anomaly. An anomaly in this dataset is denoted as either movement of non-pedestrian objects or irregular pattern of a pedestrian's movement.
Violent-flows (Hassner et al. 2012)	320x240, Varies	Behaviour Understanding	The violent flows dataset includes footage of both crowd violence and non-violence; the real-world videos were extracted from YouTube. This dataset is a benchmark to test violent/ non-violent crowd behaviour classification and recognition of violent occurrences. The shortest, longest, and the average length of the 246 videos included are 1.04, 6.52, and 3.60 seconds respectively.

Table 9: Notable benchmark datasets.

Dataset	Res. & Colour	Purpose	Description
World Expo 2010 (Zhang et al., 2016a)	720x576, RGB	Crowd segmentation and estimation	The World expo dataset collected in the Shanghai 2010 world expo is a collection of 2630 videos. The footage is annotated and captured from 245 surveillance cameras providing multiple views. 53637 crowd segmentations are manually annotated using three density level properties; high, medium and low.
WWW Crowd Attribute dataset (Shao et al. 2015)	640x360, Varies	Behaviour Understanding	The WWW dataset is one of the largest crowd datasets with 10,000 videos captured from 8,257 different environments. The authors use 94 attributes such as 'indoor', 'outdoor', 'fight', and 'band performance' to describe and annotate the videos.

3.7 Conclusion

A comprehensive overview of crowd behaviour analysis has been presented in this chapter. Crowd behaviour analysis is explored to specifically focus on understanding the behaviour of a crowd, extracting motion representations and determining occurrences of anomalous behaviour. State-of-the-art crowd anomaly detection methods have been investigated and generative adversarial networks (GANs) have been chosen as the base architecture for the framework of this research. The results achieved by different researchers utilising GANs in their framework have proven to surpass the state-of-the-art (Tables 5, 6 and 7).

The investigation into the methods for crowd anomaly detection revealed that optical flow has been used in numerous methods for temporal feature extraction. Other temporal feature extraction methods such as dynamic images have not been thoroughly considered for crowd anomaly detection despite the fact that dynamic images, in the field of action recognition, have proven to achieve better accuracy results than optical flow (Table 8). This research aims to merge dynamic images with CGANs for improved crowd anomaly detection as one of the main and novel contributions of the work.

During the experimentation stage of this research, more recent research was developed and published. The field of crowd anomaly detection is a rapidly developing field, some of the more novel work include a temporal enhanced appearance to motion generative network to model the evolution of motion and appearance of normal behaviour (Ji et al., 2020). Another noteworthy method detects anomalies based on self-supervised and multi-task learning (Georgescu et al., 2020). Ouyang and Sanchez (2020) use a deep probabilistic model that detects abnormal patterns by relying on PSNR values from their data reconstruction. Lastly, (Pourreza et al., 2021b), learn and model the interaction of normal objects using a spatio-Temporal Graph for anomaly detection.

Lastly, in this chapter the datasets used for anomaly detection were investigated and a key limitation found within these benchmark datasets is that the crowd size captured in the footage is between low to medium-density crowds. Datasets that include high-density crowds including some type of anomalous behaviour are not publicly available. Therefore, another contribution to this research is the creation and publication of a new dataset that includes both high-density crowds and anomalous behaviour (further discussed in Chapter 5).

4 Deep Generative Crowd Anomaly Detection

4.1 Introduction

The basic structure of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) is two neural networks opposing each other (adversarial). The network architecture gained popularity because of its ability to imitate data distribution. Usually, the typical neural network models are given features and a label is expected as output. The GAN model has the opposite goal; given a label the network predicts associated features. GANs generally contain two neural networks, the Generator and the Discriminator. Both networks are used to play against each other, each trying to reach its own goal. The *Generator's* goal is to learn to generate data instances that the opposing neural network, the *Discriminator*, would think is real. On the other hand, the *Discriminator* is used to learn to distinguish whether the data instances are real (from the original data/domain) or fake (generated). Other networks, discussed in Section 3.3.3, such as FCNs, SDAEs, LSTMs, and RNNs have shown successful experimental results within the crowd anomaly detection field. Moreover, GANs have been used in many fields, further discussed in Section 4.1.3, but more recently, GANs have been used for the detection of anomalies within crowds. The generative network has proved better success in the detection of anomalies than other neural networks. A simple visualisation of the GAN architecture is shown in Figure 23.

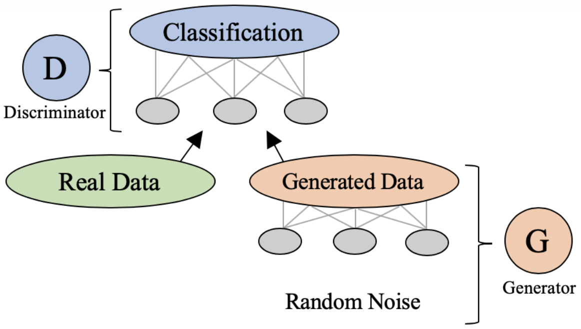


Figure 23: Simple GAN architecture. Adapted from (Hergott, 2019)

4.1.1 GAN architecture

The architecture is analogous to a two-player minimax game (Goodfellow et al., 2014). The process continues until the discriminator is more often than not fooled that the instances/samples generated by the generator are real. The discriminative model *D*, illustrated in Figure 23 above, tries to map given features to specific labels. While the generative model *G* produces new data instances to give to the discriminator to try to fool it that the generated instances are real. The models pass this data back and forth in order to strengthen their own models. In more detail, *G*s' training process aims to maximise *D*s' probability of making an inaccurate decision. In turn, *D* tries to distinguish if the data given is from the model distribution or the data distribution. Training the adversarial model step-by-step is outlined below:

1. A random distribution function is used to produce noise to be fed to the Generator *G* as the fake data.
2. The Discriminator *D* is fed both the fake data (Step 1.) and real data (training data).
3. *D* calculates *Adversarial loss* by combining the loss of real data and loss of fake data.
4. *G* imitates Step 3. by calculating its own loss of the noise data.
5. The loss variables return to their corresponding models, and the network parameters are fine-tuned with respect to the loss.
6. An optimisation method is utilised and the steps are repeated again. The number of repetitions is determined by the user.

4.1.1.1 GAN value function

The following equation is the value function of a typical GAN:

$$\underset{\mathbf{G}}{min}\,\underset{\mathbf{D}}{max}\,V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]. \tag{9}$$

$z \rightarrow$ Noise Vector

$x \rightarrow$ Training sample $\rightarrow x_{real}$

$G(z) \rightarrow$ Generator output $\rightarrow x_{fake}$

$D(x) \rightarrow$ Discriminator output for $x_{real} \rightarrow$ P(y | x_{real}) $\rightarrow [0,1]$

$D(G(z))\rightarrow$ Discriminator output for $x_{fake} \rightarrow$ P(y | x_{fake}) $\rightarrow [0,1]$

The goal of the Discriminator *D* is to **maximise** D(x) and **minimise** D(G(z)), meaning the real data is maximised and the fake data is minimised. Meanwhile, the goal of the Generator *G* is to **maximise** D(G(z)), meaning the fake data is maximised.

To calculate the loss for each network the following equations are used:

Discriminator network:

$Dloss_{real} = \log(D(x))$

$$Dloss_{fake} = \log(1 - D(G(z)))$$

$$Dloss = Dloss_{real} + Dloss_{fake} = \log((D(x)) + \log(1 - D(G(z))))$$

Generator network:

$$Gloss = \log(1 - D(G(z)))$$

Noticeably, the discriminator model is applied two times, once for the real data and the other for the fake data and the generator is applied once. A thorough description of the algorithm as detailed by Goodfellow et al. (2014) is shown below (Algorithm 1).

Algorithm 1 Minibatch stochastic gradient descent training of generative adversarial nets. The number of steps to apply to the discriminator is k. Adopted from Goodfellow et al. (2014)

```
for number of iterations do
    for $k$ steps do
        • Sample minibatch of $m$ noise samples $\{z^{(1)}, ..., z^{(m)}\}$ from noise prior $p_g(z)$
        • Sample minibatch of $m$ examples $\{x^{(1)}, ..., x^{(m)}\}$ from data generating distribution
      $P_{data}(x)$.
        • Update the discriminator by ascending its stochastic gradient:
              $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)})))]$    (10)
    end for
        • Sample minibatch of $m$ noise samples $\{z^{(1)}, ..., z^{(m)}\}$ from noise prior $p_g(z)$
        • Update the generator by descending its stochastic gradient:
              $\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \log(1 - D(G(z^{(i)})))$    (11)
end for
The gradient-based updates can use any standard gradient-based learning rule.
```

4.1.1.2 Problems in GAN

One of the major advantages of GANs is that they are very good classifiers. In comparison to CNN, they have achieved better results regarding data synthesising, and image segmentation. GANs can also handle a shortage of real data while other networks are usually data-hungry. The biggest problems in GANs are fourfold (Goodfellow, 2016; Salimans et al., 2016; Arjovsky and Bottou, 2017):

1. First, the model can go into complete or partial mode collapse. This collapse occurs when the Generator G generates a set of samples that are not diverse enough (partial) or G generates only one sample (complete).

2. The second complication that could happen is a "vanishing gradient". If the Discriminator D is trained too strongly or too weakly, the feedback given to G is not reliable, this can cause the learning process to stop.

3. Thirdly, it can be problematic finding the ideal state in which G and D are both satisfied (Nash equilibrium). Both networks feed off each other and get stronger in doing so, in this case reaching the state of Nash equilibrium is very difficult.

4. Lastly, defining the moment to stop training is hard, this is due to the lack of a true evaluation metric (Goodfellow, 2016; Salimans et al., 2016; Arjovsky and Bottou, 2017).

The enhancement of GAN training is quickly progressing and these problems have been addressed in current research producing different types of GAN architectures as discussed below.

4.1.2 Types of GAN

There are various types of generative adversarial networks (GANs) emerging. Some of the most prominent architectures are detailed below. Additionally, to illustrate the qualitative results generated from the presented methods as well as other GAN types, Appendix D includes the application of the various GANs types on the handwritten digit database (MNIST)(LeCun et al., 1998).

4.1.2.1 Deep Convolutional GAN

Deep Convolutional GANs (DCGANs) (Radford et al., 2015), is based on the standard GAN architecture. This network includes defined CNN constraints to stabilise the training process of GANs and avoid the aforementioned typical problems in GANs. Both the Generator G and the Discriminator D utilise convolutional neural networks to their own advantage. D uses a set of convolutional layers to downsample the input data with each layer. The model learns a deeper representation of the data with each layer. On the contrary, G upsamples the input data by adding noise to enlarge the input data to its original size, where downsampling reduces the sampling rate and upsampling increases the sampling rate. The training process of typical GANs usually has stability problems, and frequently the generator produces meaningless results. However, DCGANs with appropriate constraints applied to the architecture show more stability when training in diverse settings. Radford et al. (2015) applied CNN adjustments to their architecture to allow high-resolution training, as well as deeper generative models. The three adjustments applied are as follows. Convolutional net pooling functions like "max-pooling" and "average pooling" are not used, instead, strided convolutions are utilised by the generator to learn its spatial upsampling. The second adjustment applied is removing fully connected layers by using global average pooling. The last adjustment made is "Batch Normalisation" (Ioffe and Szegedy, 2015), this normalises the input to each unit in order for it to have zero mean and unit variance, achieving better stability in the learning process.

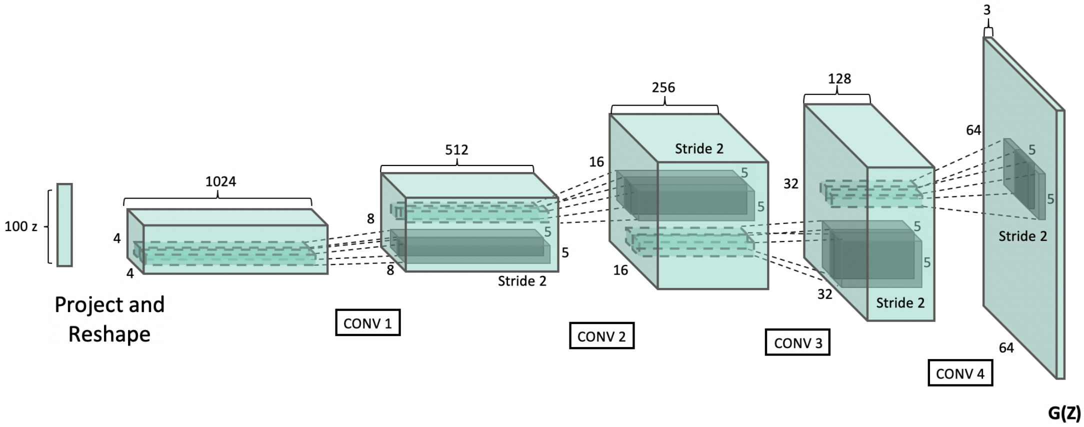


Figure 24: Deep Convolutional GAN generator. Adapted from (Radford et al., 2015)

As shown in Figure 24, the generator has four series of "four fractionally-strided convolutions" utilising batch normalisation, with the exception of the input layer. ReLU activation (Nair and Hinton, 2010) is applied in all layers, only the last output layer uses a Hyperbolic Tangent (Tanh) function. The discriminator contains four strided convolutions, similar to the generator batch normalisation is utilised for all layers with the exception of the first input layer. Unlike the generator, the discriminator works better with leaky rectified activation (Leaky ReLU) (Maas et al., 2013). The main difference between ReLU and Leaky ReLU is the former activation function takes the maximum value between the input and zero, whereas the latter activation function will allow negative values. This prevents the "dying state", where the output results given by the network are all zeros. The model is still unstable in some configurations, longer training can occasionally result in a crumble of filters subset into one oscillating mode.

4.1.2.2 Conditional Generative Adversarial Nets

Conditional Generative Adversarial Nets (CGANs) (Mirza and Osindero, 2014), are also based on the original GAN architecture. Both the G and the D are given conditional data y as an additional input, it can be any type of auxiliary data. The additional input produces higher quality data, in the application of image generation the method can control how the generated image will look. The loss function after the conditional modification is shown below:

$$\underset{\mathbf{G}}{min}\,\underset{\mathbf{D}}{max}\,V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|y)))]. \tag{12}$$

CGANs can be useful for multiple purposes such as text-to-image synthesis, image and video generation, convolutional face generation and image-to-image translation (further detailed in Section 4.2). CGANs follow the same model as DCGANs but the main difference is the conditioning vector that can control the model output. The vector should include a set of specifications as indicators of what the output should be. This data is incorporated into the learnt images as well as the input noise vector Z (Mirza and Osindero, 2014). The discriminator now evaluates both the similarities between the generated data and input data and the similarities between the generated data and the input label. The drawback of this model is the model is not purely an unsupervised method since the model requires input labels. The basic CGAN architecture is illustrated in Figure 25. As one of the main contributions of this thesis is to include image-to-image translation using CGANs (Isola et al., 2017), more details of the image-to-image translation model are documented in Section 4.2.

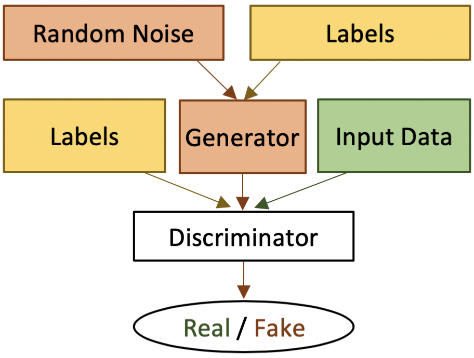


Figure 25: Basic InfoGAN architecture.

4.1.2.3 Info Generative Adversarial Nets

Info Generative Adversarial Nets (InfoGAN) (Chen et al., 2016) uses information theory in the transformation of noise into latent codes which have meaningful and systematic effects on the output of the model. The basic idea of InfoGAN is to split the input given to the generator into the standard noise and "latent code" vectors. The Mutual Information (the mutual dependence measurement between two random variables) between latent code and the generator's output is maximised to make the codes meaningful. The original GAN value function is used (Equation 9) with an additional regularisation term as shown below:

$$\underset{\mathbf{G}}{min}\,\underset{\mathbf{D}}{max}\, V_I(D,G) = V(D,G) - \lambda I(c; G(z,c)) \tag{13}$$

In this case, λ is used as a regularisation constant set to one and $\lambda I(c; G(z,c))$ is the mutual information between the two variables: latent code (c) and the output from the generator ($G(z,c)$). An illustration of the architecture is shown in Figure 26, where the Q neural network attempts to predict the latent code. Since the mutual information cannot be explicitly calculated, standard variational arguments are utilised to approximate a lower bound. To achieve this, an "auxiliary" distribution ($Q(c|x)$) is used to estimate the real $P(c|x)$. $Q(c|x)$ is modelled by a parameterised neural network and $P(c|x)$ given the generated input x indicates the likelihood of code c. The regularisation term is computed based on $P(c|x)$ not an estimate of the code c, this indicates that Q does not generate the value of code c. Q generates the statistics of the distribution and then the *likelihood* can be computed. The additional regularisation term can disentangle important data attributes to be allocated to the structure of the latent code.

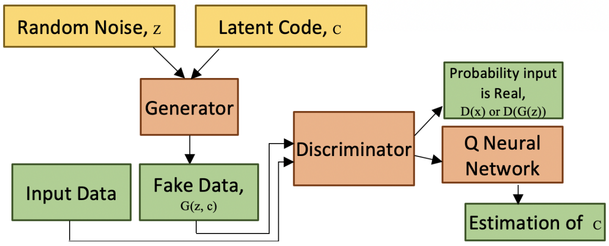


Figure 26: Basic CGAN architecture.

4.1.3 Applications of GAN

GANs have been used in many applications since they were first introduced by Goodfellow et al. (2014). For example, Radford et al. (2015) introduced deep convolutional generative adversarial networks (DCGANs) for image classification, with high accuracy results. Reed et al. (2016) also use deep convolutional GANs to synthesis images from detailed text, and Zhang et al. (2017) target the same problem using Stacked GANs. Image generation has also been implemented using GANs (Nguyen et al., 2017), the system generates high-quality images using Plug and Play Generative Networks. (Karras et al., 2017) achieved better results with image generation while decreasing the training time of the network.

GANs have also been used for image segmentation and classification. Zhu et al. (2017) employ adversarial networks to train a fully convolutional network (FCN) to detect mass in mammograms. Luc et al. (2016) combine trained convolutional semantic segmentation network with the adversarial network to segment objects from the background. On the other hand, Li et al. (2017) present a "Perceptual" GAN model to detect small objects within images. Wang et al. (2017) go in another direction with object detection, and propose a GAN network that generates "hard" samples which are images with difficult occlusions and deformations. This generated data is used to train a Fast-RCNN (FRCN) detect objects in a more robust manner. The focus of the research in this thesis is the application of GANs for crowd anomaly detection. As reviewed in Section 3.3.4, the application of GANs for this purpose has not been thoroughly investigated. Preliminary research by Ravanbakhsh et al. (2017); Liu et al. (2018b); Ravanbakhsh et al. (2019) and Vu et al. (2019) has investigated the use of GANs in crowd anomaly detection, and the resulting Equal Error Rate (EER) and Area Under Curve (AUC) produced from their frameworks have surpassed the state-of-the-art deep learning methods. This research capitalises on the aforementioned success and proposes to enhance crowd anomaly detection by extracting dynamic image representations as the temporal development features given to an image-to-image translation CGAN model.

4.2 Image-to-Image translation via CGANs

Isola et al. (2017) investigate CGANs to enhance the typical complications of image-to-image translation. Their solution named "pix2pix" can learn the input to output image mappings and a loss function to train the extracted mappings. The pix2pix method enables the utilisation of this generic method to other problems that would need a separate loss formulation. The method demonstrates its reconstruction capabilities in examples such as translation of edge maps to objects and image colourisation. The details of the pix2pix architecture based on CGANs is presented below.

Isola et al. (2017) utilise CGANs in their methods. Similar to the standard GAN, CGANs use two models to work against one another; the generator and discriminator. As noted in Section 4.1.2.2, CGANs differ from the typical GAN by using additional conditional data to guide the network into generating a specific type of image. The loss function used is the same as the standard CGAN loss function and is depicted as:

$$\mathcal{L}_{CGAN} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,y}[\log(1 - D(x, G(x,z)))] \tag{14}$$

The generator G has the objective of minimising the loss function 14 whereas the objective of the discriminator D is to maximise the loss function. To avoid unstable optimisations Isola et al. (2017) use an L1 loss function (Equation 15) as an alternative to the L2 loss function which also reduces blurring in image generation.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[||y - G(x,z)||_1] \tag{15}$$

The final objective function is noted as:

$$G^* = arg \underset{\mathbf{G}}{min} \underset{\mathbf{D}}{max} \mathcal{L}_{CGAN}(G,D) + \lambda \mathcal{L}_{L1}(G) \tag{16}$$

As noted in Isola et al. (2017), the standard encoder-decoder network presents a problem where low-level information is not shared across the network but instead the network progressively downsample the input until the bottleneck layer after which the process is reversed. To solve this issue, a “U-Net” architecture is implemented for the generator, both architectures are illustrated in Figure 27 to show their differences. The U-Net architecture includes skip connections between each encoding layer and is mirrored in the decoding layer.

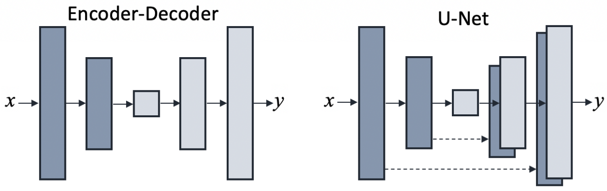


Figure 27: Left: Encoder-decoder architecture. Right: U-Net architecture. Adapted from (Isola et al., 2017).

In addition to the U-Net architecture used in the generator, Isola et al. (2017) use a Markovian discriminator named a PatchGAN. The discriminator is needed to model high-frequency structure and therefore attention is given to the structure in local image patches. PatchGAN exclusively penalises the structure at the scale of the patches, where the discriminator classifies patches as real or fake. The final output from the discriminator is based on averaging all the patch outputs of an image. For network optimisation, the network is trained to maximise $\log D(x, G(x, z))$. The architecture uses a minibatch Stochastic Gradient Descent and an Adam optimiser, the learning rate is set to 0.0002, $\lambda = 100$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. In the testing phase, the generator is run in the same way it was trained and batch normalisation utilising test batch statistics is applied. The batch size is determined based on the type of experiment.

This image-to-image translation method, pix2pix, using CGANs is been the basis of all the crowd anomaly detection methods presented in Section 4.2.1 below. Given its merits and success, pix2pix is also used as the basis of the crowd anomaly detection framework presented in this thesis as detailed in Section 4.3.

4.2.1 Anomaly detection

The application of GANs for the purpose of anomaly detection within crowds is a relatively novel approach. Although research in this area is limited, the results produced so far suggest high prospects in comparison to other deep learning models. Results from experiments conducted by the state-of-the-art GAN models used for crowd anomaly detection are presented in Table 10, the best results are indicated in bold lettering. Performance is better when Area Under Curve (AUC) increases and Equal Error Rate (EER) decreases. The methods documented in the comparison Table 10 are from research by 1. Ravanbakhsh et al. (2017), 2. Liu et al. (2018b), 3. Vu et al. (2019) and 4. Ravanbakhsh et al. (2019).

Table 10: Experimental results of CGANs for anomaly detection on UCSD

	Dataset							
	Ped-1				Ped-2			
Method	Frame Level		Pixel Level		Frame Level		Pixel Level	
	AUC(↑)	EER(↓)	AUC(↑)	EER(↓)	AUC(↑)	EER(↓)	AUC(↑)	EER(↓)
1. GAN generative	**97.4**	8	70.3	35	93.5	14	-	-
2. AnoPred	83.1	-	-	-	95.4	-	-	-
3. MLAD	82.34	23.5	66.6	**22.65**	**97.52**	**4.68**	**94.45**	**4.58**
4. GAN discriminative	96.8	**7**	70.8	34	95.5	11	-	-

Ravanbakhsh et al. (2017) and Ravanbakhsh et al. (2019) employed a conditional GAN (CGAN) model (Isola et al., 2017) to locate anomalous behaviour within crowd videos. The CGAN model generates optical flow maps after it is trained on frame pairs and the features of their corresponding optical flow (Brox et al., 2004). Anomalous behaviour is localised using two scenarios: when the error value of the generative network is high (Ravanbakhsh et al., 2017), or when the discriminator value of the CGAN model is low (Ravanbakhsh et al., 2019). The method by Liu et al. (2018b) also uses CGANs for future frame prediction to detect anomalous behaviour within a crowd. The network generates a prediction frame from a U-Net and utilises different constraints to achieve higher quality frame prediction. Some of the constraints used are adversarial training loss (Isola et al., 2017), spatial constraint based on intensity and gradient loss and an optical flow loss calculated using a pre-trained network Flownet (Dosovitskiy et al., 2015). Similar to (Ravanbakhsh et al., 2019), the research presented by Vu et al. (2019) is also based on the work of Ravanbakhsh et al. (2017). The method begins by training Denoising Autoencoders (DAEs) for each type of data: frame data and the calculated optical flow data (Brox et al., 2004). High-level features are extracted by feeding the data types to their corresponding DAE. Two CGANs are trained on a pair consisting of a frame and its corresponding optical flow high-level features previously extracted. To detect anomalous behaviour the high-level features of the testing frames and the calculated optical flow are extracted using the DAEs. Errors maps are calculated from the CGANs and a thresholding function is used to produce binary detection maps. The union of these maps is used to determine if there is an anomalous behaviour present.

The previously noted research demonstrates the capabilities of the application of CGANs to crowd behaviour anomaly detection, CGANs are also utilised in this thesis for the purposes of anomaly detection within crowds. The various applications of CGANs for anomaly detection applied in this thesis are as follows:

- The framework presented in Chong and Tay (2017), Liu et al. (2018b) and Vu et al. (2019) will be used to analyse high-density crowds as an alternative to medium-density crowd datasets commonly used. Benchmark crowd anomaly datasets currently available are limited to low and/or medium-density crowds and therefore the use of a high-density crowd dataset will exemplify the detection performance (AUC and EER) of state-of-the-art methods. Previous methods have not demonstrated the performance of their models when trained and tested on a high-density crowded environment. The state-of-the-art crowd anomaly detection methods mentioned above have been tested with the the Abnormal High-Density Crowd dataset produced in this thesis Mahmoud and Arafa (2020) and their detection performance are noted in Section 6.1.

- As an alternative to the optical flow extraction method (Brox et al., 2004) used in the research by (Ravanbakhsh et al., 2017; Ravanbakhsh et al., 2019; Vu et al., 2019), a higher performance optical flow algorithm FlowNet (Sun et al., 2017) will be applied. The framework proposed by Vu et al. (2019) is applied in conjunction with FlowNet (detailed in Section 3.4.1) to determine the effects of the advanced optical flow method to the overall detection performance (experiments are noted in Section 6.3.3). Additionally, FlowNet is applied to benchmark crowd anomaly detection datasets and the Abnormal High-Density Crowd dataset. Sample optical flow representations of these experiments are shown in Section 6.2.
- Finally, a novel approach to crowd anomaly detection is the application of Dynamic Images (Bilen et al., 2016) (detailed in Section 3.4.2) combined with image-to-image translation using CGANs (Isola et al., 2017) as an alternative to optical flow extraction. The method is tested with benchmark medium-density datasets as well as the Abnormal High-Density Crowd dataset Mahmoud and Arafa (2020). The proposed framework is detailed below and the experimental results utilising this framework on benchmark crowd datasets are presented in Section 6.3. Additionally, dynamic image extraction is applied to benchmark crowd anomaly detection datasets and the Abnormal High-Density Crowd dataset. Sample dynamic image representations of these experiments are illustrated in Section 6.2.3.

4.3 Proposed Framework

This research presents a novel approach to crowd anomaly detection which combines Dynamic Images (Bilen et al., 2016) and image-to-image translation using CGANs (Isola et al., 2017). As demonstrated in Table 10, most of the best-achieved anomaly detection results (indicated in bold) on benchmark datasets are produced using the framework proposed by Vu et al. (2019). Similar to Ravanbakhsh et al. (2017) and Vu et al. (2019), the anomaly detection method presented in this research utilises the image-to-image translation CGANs (Isola et al., 2017) to learn the transformation between frames and their corresponding image representations and vice versa based on generation loss, as noted in Section 4.2. The proposed framework builds on Vu et al. (2019)'s work and introduces dynamic images as image representations to improve crowd anomaly detection accuracy. An illustration of the proposed framework is shown below in Figure 28.

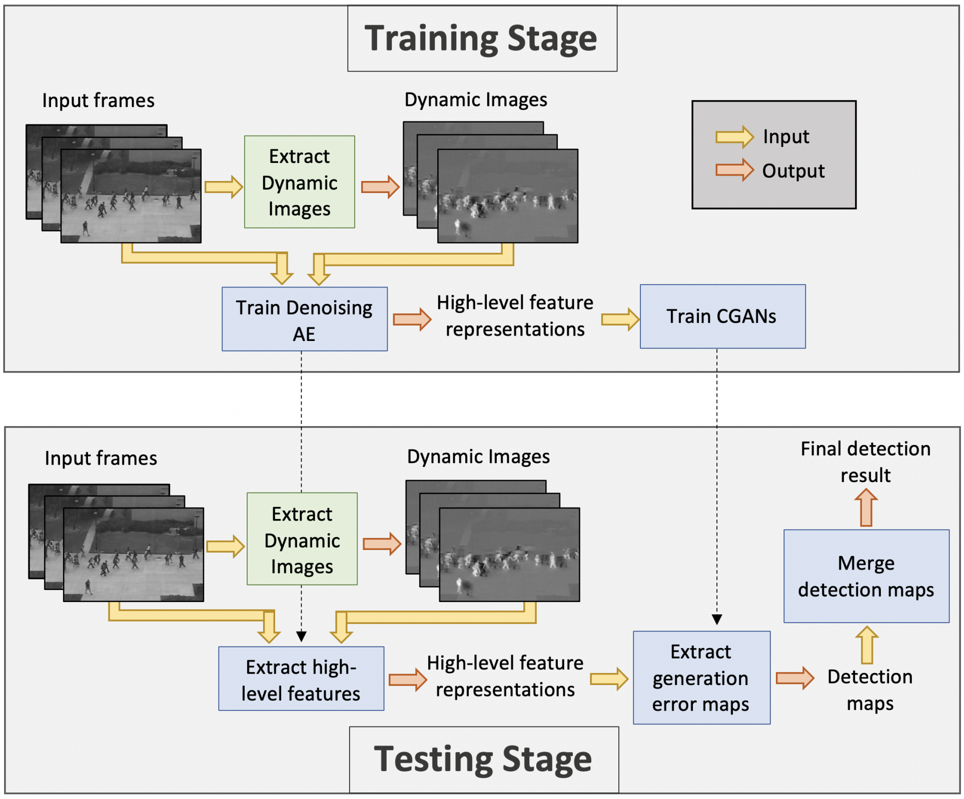


Figure 28: Illustration of the proposed crowd anomaly detection framework.

The architecture defined is divided into two main stages: training the network and testing the network (anomaly detection).

Stage 1: The training stage follows the steps below:

1. Extraction of dynamic image representations for each input frame (normal behaviour).
2. Training two different DAEs, one for the input frames and the other for dynamic images.
3. Extraction of high-level features from the input frames and dynamic images from the previously trained DAEs corresponding to its data type.
4. Training of two CGANs on the extracted high-level features of the input frames and dynamic images.

Stage 2: The testing stage follows the steps below:

1. Extraction of dynamic image representations for each input testing frame.

2. Computation of high-level features for the input frames and their corresponding dynamic image representation.

3. Computation of generation error maps using pre-trained CGANs to calculate binary detection maps for each representational level.

4. The final detection result is determined based on a merging the extracted detection maps.

4.3.1 Dynamic Image Extraction

As previously discussed in Section 3.4.2, a dynamic image is a representation of an amalgamation of input frame sequences. Following Bilen et al. (2016) and Bilen et al. (2017), a set of consecutive images (video) is represented as a *ranking function* $I_1, ..., I_T$. For each frame, I_t a feature vector representation, $\psi(I_t) \in \mathbb{R}^d$, is computed. The time average of the computed feature vectors is noted as $V_t = \frac{1}{t}\sum_{\tau=1}^{t} \psi(I_t)$ through time t. Time t is linked to a score $S(t|\mathbf{d}) = \langle \mathbf{d}, V_t \rangle$ using a *ranking function*, in this case, $\mathbf{d} \in \mathbb{R}^d$ is a vector of parameters. The parameters **d** are learned in a manner where the scores indicate the rank of the input frames so a later time correlates to a higher score. To learn **d** the RankSVM equation (Smola and Schölkopf, 2004) is used as follows.

$$\mathbf{d}^* = \rho(I_1, ..., I_T; \psi) = \underset{\mathbf{d}}{argmin}\, E(\mathbf{d}) \tag{17}$$

$$E(\mathbf{d}) = \frac{\lambda}{2}||\mathbf{d}||^2 + \frac{2}{T(T-1)} \times \sum_{q>t} max\{0, 1 - S(q|\mathbf{d}) + S(t|\mathbf{d})\} \tag{18}$$

Equation 17 is an SVM quadratic regulariser and equation 18 is a hinge-loss function that soft-counts the number of incorrectly ranked pairs $q > t$ by the scoring function. A pair is correctly ranked if there is at least a one unit margin difference between the scores meaning $S(q|\mathbf{d}) > S(t|\mathbf{d}) + 1$. The process named *rank pooling* is based on optimisation Equation(17), where a set of frames T are mapped to a single vector $\mathbf{d}^*$. *Rank pooling* is applied to RGB input frame pixels. RGB components of the frame pixels are stacked on a large vector by the operator function $\psi(T_t)$. At this stage, $\mathbf{d}^*$ is a descriptor vector containing the same number of elements as one input frame. Due to $\mathbf{d}^*$ being calculated by rank pooling it holds amalgamated information for the set of input frames and is presented as a single RGB image.

4.3.2 Training Denoising Autoencoders

Following Vu et al. (2019), Denoising Autoencoders (DAEs) are utilised to learn multilevel representations of input data. A DAE (Vincent et al., 2008) is a neural network designed to reconstruct sample data $\upsilon \in \mathcal{D}$ from the corresponding corrupted version $\tilde{\upsilon} \sim q_{noise}(\tilde{\upsilon}\ |\upsilon)$ (q_{noise} is any type of noise distribution). DAE contains two stages, an encoder and a decoder where the encoder $f_\theta(\tilde{\upsilon})$ receives an input υ to be mapped into code h in a hidden space. The decoder g_ϕ receives h and projects it back to the input space. The following objective function is used to train the network to reconstruct the original input data.

$$\underset{\theta,\phi}{min}\, \mathcal{J}_{DAE} = \underset{\theta,\phi}{min} \frac{1}{|D|} \sum ||\upsilon_i - g_\phi(f_\theta(\tilde{\upsilon}_i))||_2^2 + \gamma\Big(\sum_{l=1}^{N_e} \Big|\Big|\mathbf{W}_e^{(l)}\Big|\Big|_2^2 + \sum_{l=1}^{N_d} \Big|\Big|\mathbf{W}_d^{(l)}\Big|\Big|_2^2\Big) \tag{19}$$

f_θ and g_ϕ are deep convolutional networks of weight and bias parameters where $\theta = \{\mathbf{W}_e^{(l)}, \mathbf{b}_e^{(l)}\}_{l=1}^{N_e}$ and $\phi = \{\mathbf{W}_d^{(l)}, \mathbf{b}_d^{(l)}\}_{l=1}^{N_d}$. N_e and N_d are the numbers of hidden layers respectively corresponding to the encoder and decoder.

4.3.3 Conditional GANs

Similar to Ravanbakhsh et al. (2017); Ravanbakhsh et al. (2019); Vu et al. (2019), the proposed framework utilises image-to-image translation using CGANs (Isola et al., 2017) (detailed in Section 4.2). The generative model is used to generate an output image $G(x,z)$ from an input image x based on the learnt transformation between the two image representations. The objective function used to achieve this is as follows:

$$\mathcal{L}_{CGAN} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,y}[\log(1 - D(x, G(x,z)))] + \lambda\mathcal{L}_{L1}(x,y) \tag{20}$$

As previously noted, the generator tries to minimise the objective function and the discriminator tries to maximise it.

4.3.4 Anomaly Detection Using the Proposed Framework

To detect anomalies using the proposed framework (Figure 28), the method is divided into two stages, training and a testing stage.

4.3.4.1 Training

Initially, the training stage begins with computing the dynamic image representations of the input video (normal behaviour). Following the approach detailed in Section 4.3.1, dynamic images $\mathbf{d}_t^*$ are extracted for every 10 consecutive frames $(F_t, ..., F_{t+10})$ of the set of input frames. The dynamic image for each frame is a summarisation of the motion data for the next 10 frames. Then two DAEs; DAE$_F$ and DAE$_{\mathbf{d}^*}$, are trained on the input frames and their corresponding dynamic image respectively. The same number of layers are used for both DAEs, following Vu et al. (2019) the encoder contains convolutional layers with stride = 2 and kernel size = 5 x 5. This is followed with batch normalisation layers as well as leaky ReLU activation functions. Similar to the encoder, the decoder follows the same architecture but the convolutional layers are replaced by deconvolutional layers. Both DAEs are trained using Adagrad optimisation function, $\gamma = 1$, learning rate = 0.1 and the networks are trained for 500 epochs. Each frame F_t and dynamic image $\mathbf{d}^*_t$ is fed into its corresponding pre-trained DAE to achieve the activations of each at every encoding layer. The activations are then normalised to zero-mean and unit-variance and clipped to [-1,1] to compute $F^{(k)}_t$ and $\mathbf{d}^{*(k)}_t$ as the abstract representation in the k^{th} level of the input frame data and dynamic image data respectively. k denotes a number between $0 \leq k \leq N_e$ (the number of hidden layers in the DAEs). For a set of input frames, the abstract representation is noted as $D_F = \{F_t\}^{N_f}_{t=1}$ where N_f is the number of frames in the input set. Therefore, the previous step has calculated $D^{(k)}_F = \{F^{(k)}_t\}$ and $D^{(k)}_{\mathbf{d}^*} = \{\mathbf{d}^{*(k)}_t\}$ for the frame data and dynamic image data respectively to be given to the CGANs.

Similar to (Ravanbakhsh et al., 2017) and (Vu et al., 2019) two CGANs are trained on all levels of representation k. The generator $G^{(k)}_{\mathbf{d}^* \to F}$ is trained to generate the frame image $F^{(k)}_t$ from the dynamic image representation $\mathbf{d}^{*(k)}_t$ and the corresponding discriminator is trained with input $D^{(k)}_{\mathbf{d}^*}$ and label $D^{(k)}_F$. The second generator $G^{(k)}_{F \to \mathbf{d}^*}$ is trained to generate the dynamic image representation $\mathbf{d}^{*(k)}_t$ from the frame image representation input $F^{(k)}_t$ and its corresponding discriminator is trained with input $D^{(k)}_F$ and label $D^{(k)}_{\mathbf{d}^*}$. The training settings are the learning rate = 0.0002, $\lambda = 100$ and batch size = 1. After the CGANs have been trained the output is N_e number of $G^{(k)}_{\mathbf{d}^* \to F}$ and $G^{(k)}_{F \to \mathbf{d}^*}$ at all abstract level representations k that are used to detect anomalies (test).

4.3.4.2 Testing

The testing stage of the method takes an input of image frames F_t and calculate their corresponding dynamic image representations $\mathbf{d}^*_t$. The pre-trained DAEs; DAE$_F$ and DAE$_{\mathbf{d}^*}$ are utilised to extract the high-level feature representations $F^{(k)}_t$ and $\mathbf{d}^{*(k)}_t$ from the corresponding inputs F_t and $\mathbf{d}^*_t$. The pre-trained CGANs take the previously computed high-level representation as input for each representation level k to generate a frame image $\hat{F}^{(k)}_t = G^{(k)}_{\mathbf{d}^* \to F}\big(\mathbf{d}^{*(k)}_t, z\big)$ and dynamic image representation $\hat{\mathbf{d}}^{*(k)}_t = G^{(k)}_{F \to \mathbf{d}^*}\big(F^{(k)}_t, z\big)$. The value of $F^{(k)}_t$, $\hat{F}^{(k)}_t$, $\mathbf{d}^{*(k)}_t$ and $\hat{\mathbf{d}}^{*(k)}_t$ are set to 0 at locations where the dynamic image presents no motion. This is based on the premise that an anomaly occurs in regions that contain motion, this also helps with the anomaly detection speed.

Generation error maps are then calculated based on the difference between the generated features and the original features denoted as $e^{(k)}_{F,t} = F^{(k)}_t - \hat{F}^{(k)}_t$ and $e_{\mathbf{d}^*,t} = \mathbf{d}^{*(k)}_t - \hat{\mathbf{d}}^{*(k)}_t$. The generation error maps extracted are normalised into [0, 1] for every channel as follows:

$$\bar{e}^{(k)}_{F,t} = \Big[e^{(k)}_{F,t,j}/m_{F,j}\Big]^{N^{(k)}_F}_{j=1} \tag{21}$$

$$\bar{e}^{(k)}_{\mathbf{d}^*,t} = \left[e^{(k)}_{\mathbf{d}^*,t,j} / m_{\mathbf{d}^*,j} \right]_{j=1}^{N^{(k)}_{\mathbf{d}^*}} \tag{22}$$

$N^{(k)}_F$ and $N^{(k)}_{\mathbf{d}^*}$ are the numbers of channels of the generation error maps and $m_{F,j} = max_{t,x,y} e^{(k)}_{F,t,j}(x, y)$ and $m_{\mathbf{d}^*,j} = max_{t,x,y} e^{(k)}_{\mathbf{d}^*,t,j}(x, y)$ are the maximum errors in all locations of the set of input frames for the j^{th} channel. A summation of the normalised generation error maps is computed as follows, $\bar{e}^{(k)}_t = \bar{e}^{(k)}_{F,t} + \alpha \bar{e}^{(k)}_{\mathbf{d}^*,t}$. Following (Ravanbakhsh et al., 2017; Ravanbakhsh et al., 2019), we set $\alpha = 2$ to control the effect of each type of feature. The combined generation error maps for the set of input frames are noted as $E^{(k)} = \{\bar{e}^{(k)}_t\}$ and then consecutive frames are using a sliding frame window = 5 to smooth the generation error maps. Determining an anomaly is based on a comparison between $E^{(k)}$ and a predetermined threshold β, where if $\bar{e}^{(k)}_t(x, y) > \beta$, (x, y) being the pixel location on the k^{th} frames, then the binary detection map $D^{(k)}_t(x, y) = 1$. If $\bar{e}^{(k)}_t(x, y) \leq \beta$ then the binary detection map $D^{(k)}_t(x, y) = 0$, with value of 1 indicating an anomalous pixel and value of 0 indicating a normal pixel.

Finally, to apply multilevel anomaly detection the extracted detection maps are combined using the algorithm by Vu et al. (2019). Combining detection maps from different levels supports and enhances incorrect detections. The detection maps are combined to consolidate the detected anomalous objects over different levels.

4.4 Conclusion

Generative Adversarial Networks (GANs) have shown strong promise in the field of crowd behaviour anomaly detection. In comparison to other deep models, Conditional GANs (CGANs) particularly have demonstrated effective capabilities detecting anomalies in crowd behaviour. The image-to-image translation research by Isola et al. (2017) has been the base of multiple architectures (Ravanbakhsh et al., 2017), (Liu et al., 2018b), (Ravanbakhsh et al., 2019) and(Vu et al., 2019) for crowd anomaly detection.

These methods have shown promising anomaly detection results compared to other state-of-the-art crowd anomaly detection methods. Consequently, the basis of the proposed crowd anomaly detection framework uses image-to-image translation using CGANs. The novelty of the proposed framework is combing image-to-image translation using CGANs and Dynamic Images as motion representation. The novel framework is introduced as one of the main contributions of this thesis. As previously documented in Section 3.4 dynamic image representations have been used in the action recognition field and the experimental results exhibit the benefits of their application. Therefore, the proposed method is expected to enhance performance results (AUC and EER) compared to state-of-the-art methods for the detection and localisation of anomalies within crowds. The details of the training and testing process of the proposed framework were documented in this chapter and the experimental results produced from these applications are presented in Chapter 6. The details of the high-density created, Abnormal High-Density Crowd dataset Mahmoud and Arafa (2020), are documented in the next chapter.

5 Abnormal High-Density Crowd Dataset

The availability of a benchmark datasets containing high-density crowd footage is very limited. Datasets such as the Avenue (Lu et al., 2013), UCSD (Chan et al., 2008) and UMN (University of Minnesota, 2006) datasets are examples of this. Furthermore, a dataset of high-density crowd footage that includes anomalous behaviours such as stampedes, overcrowding, violence or panic is not available. Since available datasets are inadequate to provide these features, a new dataset containing scenes which adheres to these constraints was created and published on Kaggle (Mahmoud, 2019). This dataset is a compilation of public footage collected from various online resources containing scenes of anomalous crowd behaviour. To practically evaluate state-of-the-art crowd anomaly detection methods in a high-density crowd environment, the methods discussed in Section 3.3 were tested on this new dataset and the results are presented in Chapter 6.

Prior to the creation of this dataset, simulation software/methods of high-density crowds were used investigated (multiple software were used to simulate human shaped models as a part of a crowd) to generate the dataset. State-of-the-art crowd and pedestrian simulation software were chosen based on the quality of a sample simulation, user friendliness and pricing. Initial investigation of the state-of-the-art software/methods to simulate highly dense crowds showed promise, the human models were created and placed in an area as part of a crowd. The software was used to simulate a high-density crowd exiting a room. The majority of the methods utilise agent-based modelling techniques that produce videos of crowds walking, running, looking, stopping, changing direction and avoiding collisions. However, these modelling techniques are based on prior knowledge and do not incorporate crowd emotion features (Zhao et al., 2018). Anomalies such as fights, violence, stampedes, riots and panicking behaviours require the extraction and simulation of emotional features to be incorporated within the crowd modelling techniques. These psychological features enable the crowd to perform reactive behaviour such as frantic dispersion, pushing, hiding, fighting, etc (Dickinson et al., 2019) and (Dupre and Argyriou, 2019). For these reasons, the use of software to create and label an abnormal high-density crowd dataset was ceased.

Due to the limitations of the state-of-the-art methods in crowd modelling for simulation, the collection of real-world crowd footage containing anomalous events was favoured. The new dataset is named "Abnormal High-Density Crowd Dataset" (AHDCrowd) (Mahmoud, 2019). Details of the data collection process, privacy issues, pre-processing and annotations for this dataset are documented below. A detailed description of the dataset including the current dataset statistics are also presented. Lastly, the usage, evaluation protocols, challenges and limitations are discussed.

5.1 Data Collection

The collection process involved an internet search of keywords such as: "crowd fight", "crowd violence", "crowd stampede", "riots", and "violent mobs". For the purpose of this research, these keywords were used to find footage of abnormal behaviour within high-density crowds. More specifically, the focus of this research is the prompt detection of crowd abnormal behaviour to avoid chaotic and possibly hazardous events. These keywords were used to find events that have demonstrated crowd abnormalities which have led to disorderly behaviours. A total of 13 video sequences from 4 events were collected based on the challenges and limitations experienced while collecting these data. Some of these challenges are the scarcity of high-density crowd footage and privacy issues, more challenges and limitations are described in Section 5.4. The videos vary in resolution, view angle and length, each of which is documented below. All videos were downloaded from YouTube, privacy issues are discussed below, and the veracity of the annotations has been established. All scenes contain high-density crowds in a public outdoor environment. Sample images of each scene are illustrated for each event, the anomalies are not acted out they are based on actual occurrences.

5.1.1 Privacy

The privacy of the individuals captured in the collected footage is addressed using the YouTube privacy policy for identity protection (YouTube, 2020). Individual privacy is violated when specific guidelines set by YouTube are breached. If an individual can be uniquely identified from the footage through any of the below factors, then privacy has been violated (YouTube, 2020):

- Image or voice
- Full name
- Financial information
- Contact information
- Other personally identifiable information

The collected footage adheres to these guidelines, moreover, the collected footage remains available online indicating the footage has not been flagged or removed due to privacy violations.

5.1.2 Pre-processing

After collecting the footage from YouTube some pre-processing was applied to structure the videos in a suitable and user-friendly dataset. All footage is pre-processed using lossless techniques following the 4 steps below:

1. The exact moments of anomaly occurrences are determined (based on personal observation) based on segment and frame levels.
2. Video footage is trimmed to focus on the occurrences of normal behaviour and anomalous behaviour identified in the previous step.
3. Frames are extracted from the trimmed footage and divided into training (normal behaviour) and testing (abnormal behaviour) sets.
4. Where necessary, the extracted frames from the footage are cropped to place focus on the crowded scenes and less focus on the background.
   (a) Frames extracted from some videos have been cropped to remove some of the background scenes. A sample of this cropping is illustrated in Figure 29.
   (b) This is done to reduce excess or indirect computation time when utilised in the application of processing methods.
   (c) Cropping has been sufficiently applied to prevent bias training and reduce overhead.
5. Extracted footage frames are compressed to reduce storage.

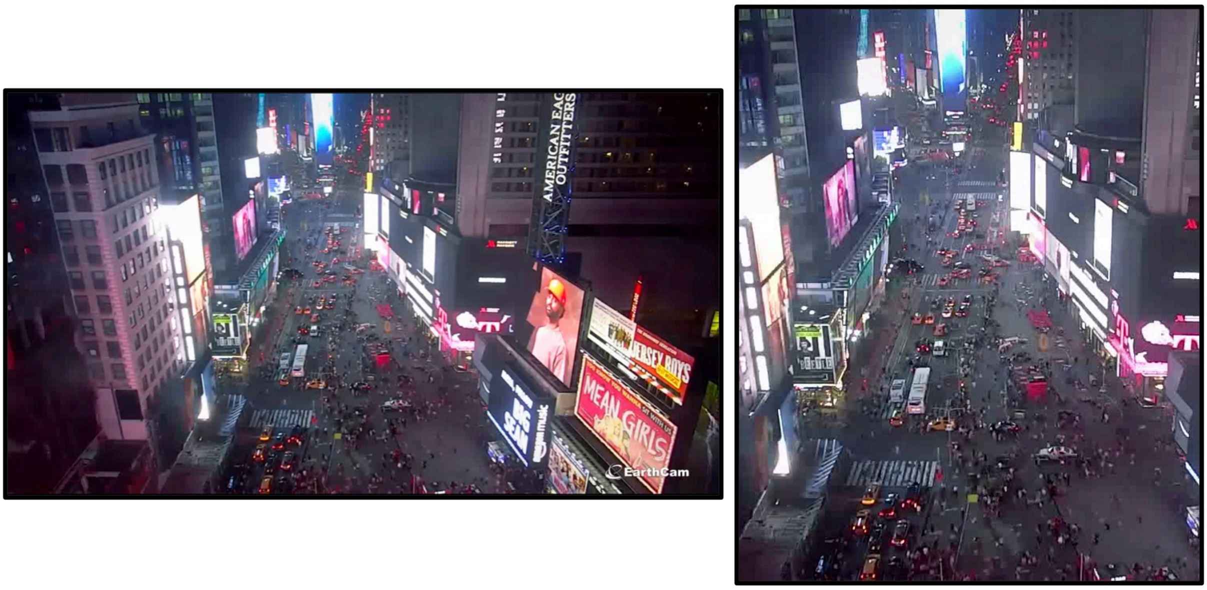


Figure 29: Example of cropping applied to frames of specific videos (Mahmoud, 2019). Left: sample original frame, Right: sample cropped frame (zoomed in).

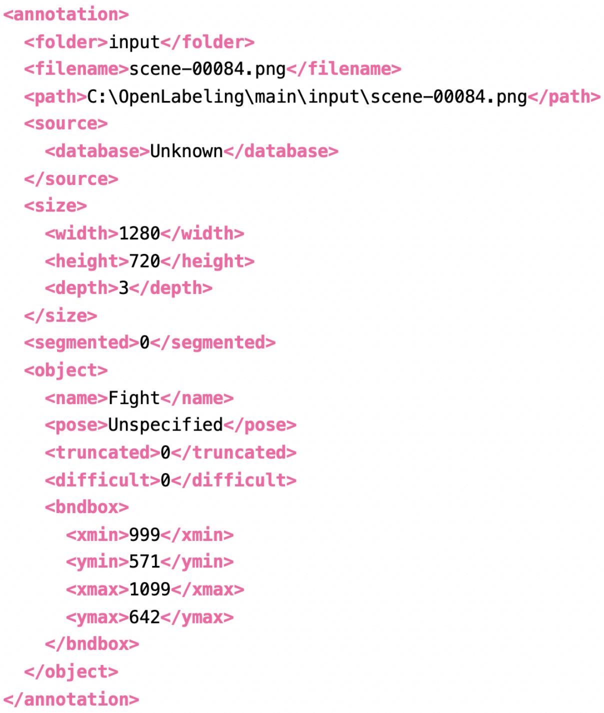
5.1.3 Annotation

While there are various automated dataset annotation tools for a computer vision task such as object recognition, automated annotation tools for anomaly detection are not available. The annotations tools for object recognition have been trained with pre-defined feature detection algorithms to identify, categorise, and annotate these objects. However, as previously discussed, anomalies are not conformed to a set of features. They are more diverse and cannot be easily defined. Hence, personal-observation (the author of this research) was used instead to precisely detect and annotate the occurrences of an anomalous event down to the second the event occurred. Two types of annotations are identified: Segment-level and Frame-level annotations.

1. Segment-level annotations:
   Annotations of the specific timing of when an anomalous event has occurred to when it ended have been documented for each collected video. Annotations of this kind are named segment-level annotations, where segments of the video (containing anomalies) are annotated as anomalous, this has been the type of annotating applied to three of the four anomalous events collected for this dataset. The fourth scene is annotated using Frame-level annotations detailed below.

2. Frame-level annotations:
   The second type of annotating has been applied to the Love Parade incident footage (in Section 5.2 below), this is a frame-level annotating method. The start and end of the abnormal behaviour segments in the video are determined and annotated as such. Specific frames that contain anomalous the events "Crowd Surge" or "Fight" are labelled, and the specific location (within each frame) of these anomalies are also annotated.

The LabelImg (Tzutalin, 2017) software was used to manually label objects (anomalies for the purposes of this research) with a bounding box for each frame within a video. The Frame-level annotations are saved as an XML or text files indicating the location and label for each event which is either a "Crowd Surge" or a "Fight". The file name of the produced annotation file is identical to the file name of the corresponding frame. The XML file contains a set of details about the labelled frames, as shown in Figure 30, some of the prominent details include the folder name, filename and path of the input frame, additionally, the width, height and depth of the frame are also included. Finally, the name of the label ("Fight" or "Crowd Surge"), and the location of the bounding box (x, y, height, width) surrounding the anomaly for the corresponding frame are documented.

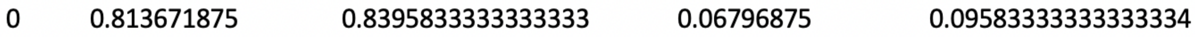


```
<annotation>
  <folder>input</folder>
  <filename>scene-00084.png</filename>
  <path>C:\OpenLabeling\main\input\scene-00084.png</path>
  <source>
    <database>Unknown</database>
  </source>
  <size>
    <width>1280</width>
    <height>720</height>
    <depth>3</depth>
  </size>
  <segmented>0</segmented>
  <object>
    <name>Fight</name>
    <pose>Unspecified</pose>
    <truncated>0</truncated>
    <difficult>0</difficult>
    <bndbox>
      <xmin>999</xmin>
      <ymin>571</ymin>
      <xmax>1099</xmax>
      <ymax>642</ymax>
    </bndbox>
  </object>
</annotation>
```

Figure 30: Sample XML labelling file of the 84^{th} video frame of the Pride parade test footage

The TXT format of the labelled files is designed to contain only the essential data about the frames. Five values are documented for each frame, as shown in Figure 31, as follows.

- The first value is an integer representation of the label ("Fight" = 0 and "Crowd Surge" = 1).
- The second and third value are the x and y locations of the anomalous bounding box relative to the size of the frame.
- Finally, the fourth and fifth values are the height and width of the anomalous bounding box relative to the size of the frame.



```
0	0.813671875	0.8395833333333333	0.06796875	0.09583333333333334
```

Figure 31: Sample TXT labelling file of the 90^{th} video frame of the Pride parade test footage

The annotations were determined and annotated using personal observation of when and where an anomalous event has happened. The videos were methodically viewed to find the specific timestamps at which an anomaly has started and ended. After the timestamps have been determined, a specific start and end frame within that time-frame are selected as the start/end of an anomaly on the frame-level. The extracted frames between the start and end of an anomaly are placed in their corresponding folders (train or test).

5.1.3.1 Veracity of annotations

To further confirm the veracity of the annotations noted for each video of the dataset, other computing researchers were asked to determine and annotate the videos. Only two researchers were available and they were recruited to confirm the occurrences (and in some cases the location) of anomalies. The researchers were asked to specify a time (minute and second) for when an anomaly (and in some cases the type) has occurred based on their personal observation. This was done to prevent possible inaccuracies or biased annotations of the footage. Below is a description of the tasks they were given.

- Document the specific start and end time for any anomalous occurrences in each video of the four scenes in the dataset.
- Additionally, document the specific start and end time as well as the type of anomaly ("Crowd Surge" or "Fight") for the third scene specifically (Love Parade).

To verify the consistency of the results produced by the researchers against the original annotations, Start and End timestamps were compared and the annotations were validated as true if both timestamps were similar for all researchers allowing an error margin of +/- one second. After applying these constraints to the results from both researchers it has been confirmed the originally extracted anomalous segments for each video from each scene is correct. A confirmation of the locations and types of anomalous incidents in the third scene were also validated to be in alignment with the suggested locations and types by the researchers. This was validated by comparing the distance between the researchers suggested locations to the originally noted locations; if the suggested locations were in the same region (with a margin of error equivalent to +/- 1 cm within the localised frame) as the original location then it is considered as true localisation.

5.1.4 Summary Description

The dataset consists of 4 scene incidents named: 1) Times Square, 2) Las Vegas, 3) Love Parade and 4) Italy. These were the only anomalous videos available based on the data collection process described in Section 5.1. Each incident is detailed below.

1. Times Square: Times Square frantic dispersion from Three Angles were available, so they were used as different scenes[1]. The footage of the three angles is concatenated into one video showing a quick dispersion of a highly dense crowd instigated by a motorcycle backfire. The crowd thought they heard gunshots and started to panic. The footage is divided into three viewpoints, where each viewpoint is divided into training (normal) and testing (abnormal) frames. Video is 29.97 FPS.

[1] https://www.youtube.com/watch?v=5g3XOuzFCSM

(a) View 1: Footage is shot from an angled view. Normal behaviour is shown at 00:00:00-00:00:12, while abnormal behaviour is shown at 00:00:12-00:00:47 of the video. Abnormality begins at the top-right corner of the image. The dimensions of the frames are 1280x720. Segment-level labelling: Train folder contains 379 "normal" frames and Test folder contains 1026 frames in total, where frames 0 - 100 are "normal" frames and frames 101 - 1026 are "abnormal" frames.

(b) View 2: Footage is shot from a closeup almost eye-level shot. Normal behaviour is shown at 00:00:48-00:00:52 and abnormal behaviour is shown at 00:00:53-00:01:32 of the video. Abnormality begins at the right side of the image. The dimensions of the frames are 1280x720. Segment-level labelling: Train folder contains 150 "normal" frames and Test folder contains 1173 frames in total, where frames 0 - 30 are "normal" frames and frames 31 - 1173 are "abnormal" frames.

(c) View 3: Footage is shot from another angled shot. Normal behaviour is shown at 00:01:33-00:01:39, and abnormal behaviour is shown at 00:01:39-00:02:18 of the video. Abnormality begins at the mid-region (closer to the right side) of the image. The dimensions of the frames cropped down to 580x720 to place more focus on the crowd. Segment-level labelling: Train folder contains 184 "normal" frames and Test folder contains 1151 frames in total, where frames 0 - 30 are "normal" frames and frames 31 - 1151 are "abnormal" frames.

Sample frames are shown in Figure 32.

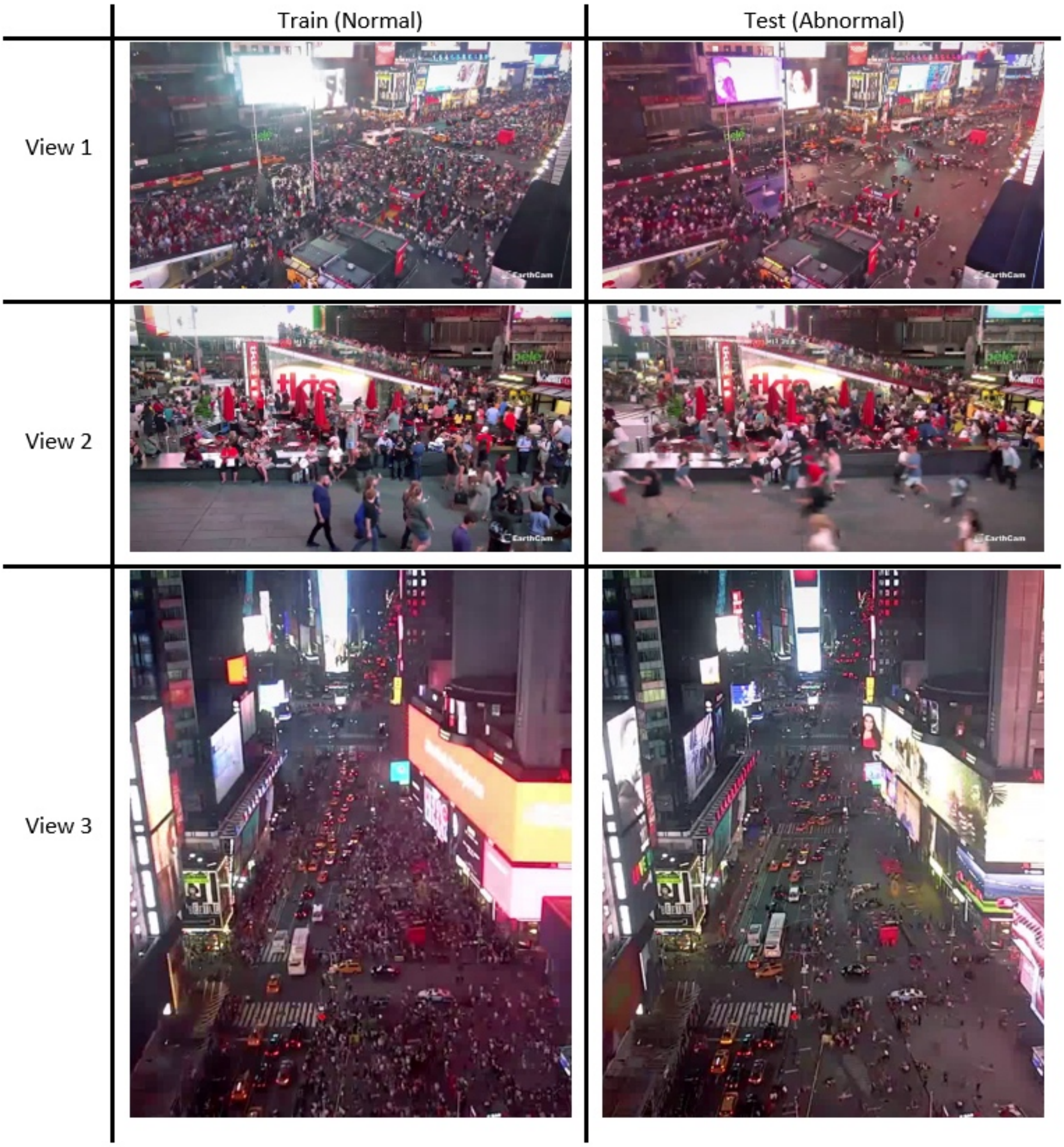


Figure 32: Sample images from each view angle of the Times Square incident (Mahmoud, 2019).

2. Las Vegas: Las Vegas Mass Shooting CCTV Video from Mandalay Bay Hotel Roof[2]. The footage shows rapid scattering within the crowd, people hiding, and people falling down. The footage is divided into training (normal) and testing (abnormal) frames. Video is 15.17 FPS. Segment-level labelling: Train folder contains 4347 "normal" frames and Test folders contain a total of 7244 frames in total, where frames 0 - 30 are "normal" frames and frames 31 - 7244 are "abnormal" frames.

[2] https://www.youtube.com/watch?v=9LHdda45k18

(a) Train: Footage is shot from a wide-angled shot. Normal behaviour is present between 00:11:17-00:16:05 in the original video. Dimensions are cropped to 992x468 to place more focus on the crowd.

(b) Test 1: Footage taken at the same angle as the training footage. Abnormal behaviour is present between 00:16:06-00:17:17 in the original video. Dimensions are cropped to 992x468 to place more focus on the crowd.

(c) Test 2: Footage taken at a closer angle. Abnormal behaviour is present between 00:17:23-00:18:23 in the original video. Dimensions of frames are 1280x720.

(d) Test 3: Footage taken at a very close angle and in greyscale format. Abnormal behaviour is present between 00:19:08-00:21:12 in the original video. Dimensions of frames are 1280x720.

(e) Test 4: Footage taken at a very close angle and in greyscale format. Abnormal behaviour is present between 00:21:15-00:25:01 in the original video. Dimensions of frames are 1280x720.

Sample frames are shown in Figure 33.

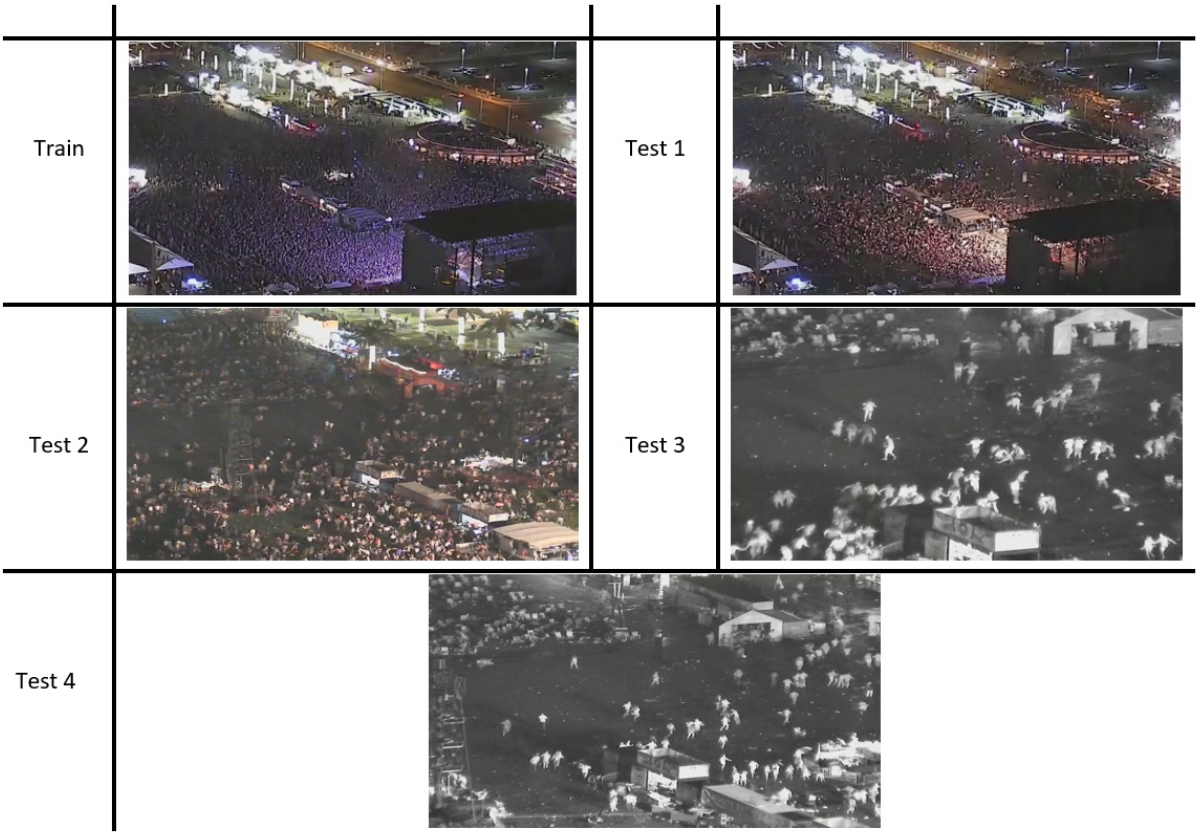


Figure 33: Sample images from each angle of the Las Vegas shooting incident (Mahmoud, 2019).

3. Love Parade: Love Parade disaster[3]. The footage shows occurrences of over-crowding, crowd surges and a fight in the footage of the 2010 Love Parade. The footage is divided into training (normal) and testing (abnormal) frames. Video is 25 FPS. Dimensions of frames are all 1280x720. All footage is shot at an angled viewpoint.

   (a) Train 1: Normal behaviour is present between 00:10:43-00:10:47 in the original video.

   (b) Train 2: Normal behaviour is present between 00:11:04-00:11:07 in the original video.

   (c) Test: Abnormal behaviour is present between 00:10:48-00:11:03 in the original video. Frame-level labelling for anomalies is available and saved in XML format. The anomalies in this scene are annotated as either "Crowd Surge" of "Fight".

   Sample frames are shown in Figure 34.

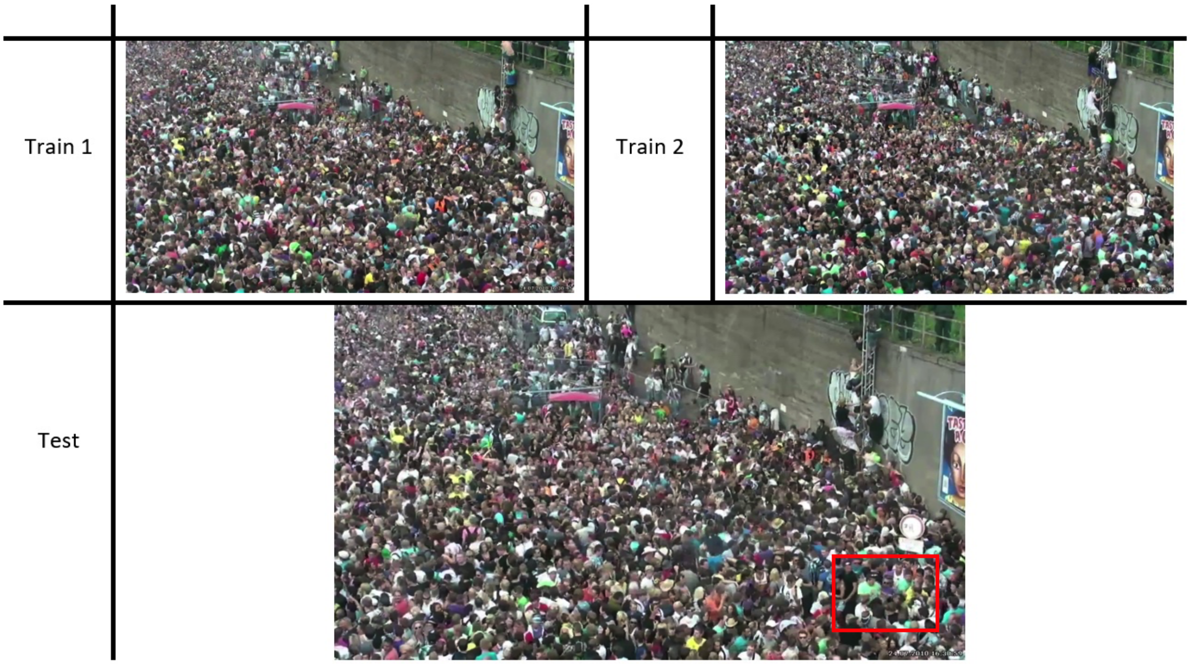


Figure 34: Sample images from the Love Parade incident, the anomaly is located using a red bounding box (Mahmoud, 2019).

4. Italy: Juventus fans panic and rapidly disperse after bomb a scare[4][5]. The footage only captures when the crowd has started to quickly disperse, and hence prevented the extraction of training or "normal" data. The footage is divided into two viewpoints, each

[3] https://www.youtube.com/watch?v=QpzISdBE3dA&t=1s
[4] https://www.youtube.com/watch?v=IP9wACjt8MU
[5] https://www.youtube.com/watch?v=yuqcNgcgzIA

viewpoint contains only testing (abnormal) frames. Video is 25 FPS.

(a) View 1: footage is shot from a wide, close and eye-level angle. Abnormal behaviour is present between 00:00:00-00:00:28 of the video. Abnormality begins at the right side of the image. Dimensions of the frames are 1280x720. The test folder contains 702 frames in total, where frames 0 - 702 are "abnormal".

(b) View 2: footage is shot from a wide-angle. Abnormal behaviour is present between 00:00:00-00:00:28 of the video. Abnormality begins at the mid-left side of the image. Dimensions of the frames are 880x720. The test folder contains 702 frames in total, where frames 0 - 702 are "abnormal".

Sample frames are shown in Figure 35.

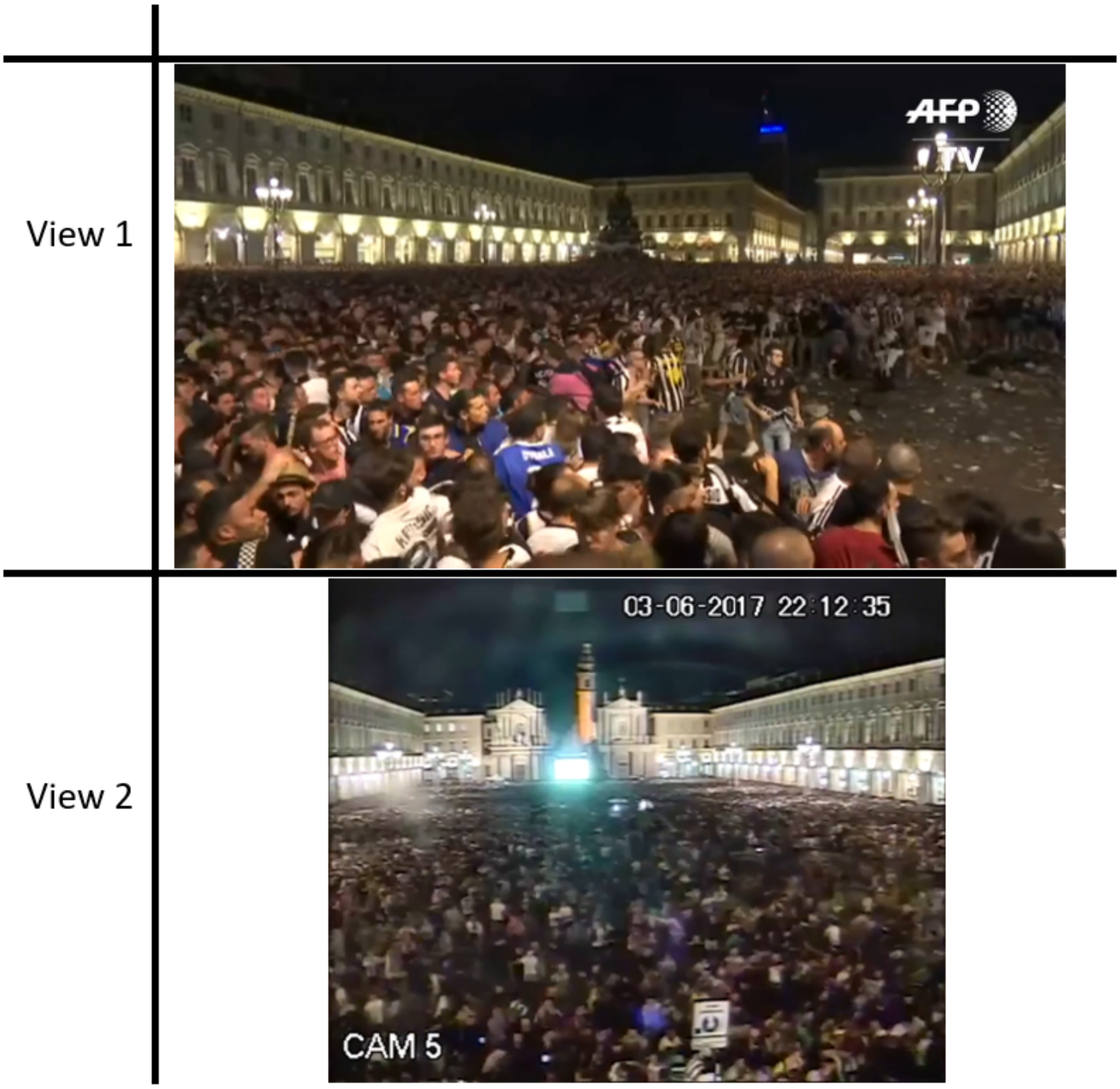


Figure 35: Sample images from the two angles of the Italy bomb scare incident (Mahmoud, 2019).

5.2 Dataset Description

An illustration of the Abnormal High-Density Crowd dataset structure is shown in Figure 36. The dataset is divided into several folders and files each of which is also divided into more folders and files. A detailed description of the folders and files are noted below:

1. File: "Dataset Image.png": a combination of sample images for each scene, this helps clarify what the dataset footage contains without the need to download.
2. Folder: "Times Square": this folder contains the Times Square incident footage divided into three different views.
   (a) File: "Footage.avi": this video file is the captured footage of the entire incident from the three angles consecutively.
   (b) Folder: "View_1:" this folder includes the training and testing frames of the first view, it is captured from a high angle view of the street.
      i. Folder "Train": this folder contains 379 extracted frames of the captured crowd. At this point, the crowd is in a "normal" state.
      ii. Folder "Test": this folder contains the remaining 1026 extracted frames for this view of the incident. At this point, the crowd starts in a "normal" state then (at frame $\sim=$ 100) beings dispersing erratically for the remaining frames, which is considered as "abnormal".
   (c) Folder "View_2:" this folder includes the training and testing frames of the second view of the incident, it is captured from a close eye-level angle view of the street.
      i. Folder "Train": this folder contains 150 extracted frames of the captured crowd. At this point, the crowd is in a "normal" state.
      ii. Folder "Test": this folder contains the remaining 1173 extracted frames for this view of the incident. At this point, the crowd starts in a "normal" state then (at frame $\sim=$ 30) beings dispersing erratically for the remaining frames, which is considered as "abnormal".
   (d) Folder "View_3:" this folder includes the training and testing frames of the third view of the incident, it is captured from a remote and high straight angle view of the street.
      i. Folder "Train": this folder contains 184 extracted frames of the captured crowd. At this point, the crowd is in a "normal" state.
      ii. Folder "Test": this folder contains the remaining 1151 extracted frames for this view of the incident. At this point, the crowd starts in a "normal" state

then (at frame $\sim=30$) beings dispersing erratically for the remaining frames, which is considered as "abnormal".

3. Folder "Las Vegas": this folder contains the Las Vegas incident footage divided into a train folder and four test folders. When the incident occurred, CCTV operators progressed to zoom in into the crowd, leading to four views that differ in the closeness of the shot but captured from the same camera.

   (a) Folder "Train": This folder contains the training footage of the incident in two formats; video format and the corresponding extracted frames.

      i. File "Footage.mp4": this video file is the captured footage of the "normal" incident and the extracted frames are below. The footage is captured from a high angled view.

      ii. Files: the remaining files are the 4347 extracted frames; the crowds are enjoying the concert in a "normal" state.

   (b) Folder "Test_1": This folder contains the first testing footage of the incident in two formats; video format and the corresponding extracted frames.

      i. File "Footage.mp4": this video file is the captured footage of the incident, the video begins with the audience in "normal" state similar to the training data, then when the gunshots were noticed the audience started to disperse quickly towards the exits. This is considered as the "abnormal" state.

      ii. Files: the remaining files are the 1063 extracted frames, the crowds are enjoying the concert until the "abnormal" state begins (at frame $\sim=30$). The view angle of this test data is the same as the training data.

   (c) Folder "Test_2": This folder contains the second testing footage of the incident in two formats; video format and the corresponding extracted frames.

      i. File "Footage.mp4": this video is a continuation of the previous video after the CCTV operator has zoomed in into the crowd.

      ii. Files: the remaining files are the 920 extracted frames; all these frames are "abnormal".

   (d) Folder "Test_3": This folder contains the third testing footage of the incident in two formats; video format and the corresponding extracted frames.

      i. File "Footage.mp4": this video is a continuation of the previous video after the CCTV operator has zoomed in further into the crowd and switched to grayscale capturing.

ii. Files: the remaining files are the 1854 extracted frames, all these frames are "abnormal".

(e) Folder "Test_4": This folder contains the fourth testing footage of the incident in two formats; video format and the corresponding extracted frames.

i. File "Footage.mp4": this video is a continuation of the previous video after the CCTV operator has zoomed out from the crowd but continued to with grayscale capturing.

ii. Files: the remaining files are the 3407 extracted frames; all these frames are "abnormal".

4. Folder "Love Parade": this folder contains footage of the love parade incident where instances of "Fight" and "Crowd Surge" occur. The footage is divided into two training folders and one test folder. The training footage is captured before and after the anomalous incidences.

(a) File "Footage.mp4": this video file is the captured footage of the entire scene from the beginning where the crowd was in a "normal" state then the "abnormal" state occurs then the scene goes back to a "normal" state.

(b) Folder "Train_1": this folder contains 131 frames of a highly dense crowd gathered in a public area; this is considered as the "normal" state of the crowd.

(c) Folder "Train_2": this folder contains 119 frames of the same crowd after the two anomalous incidents have occurred and the crowd has returned to a "normal" state.

(d) Folder "Test": this folder contains the "abnormal" state frames and their corresponding frame-level labels, the labels are saved using two formats for user convenience.

i. Folder "Labels": this folder contains the labelling files for each extracted frame.

A. Files "XML labels": the XML version of the labels contain a set of details about the frames including the name of the label, the location of the incident (bounding box) and more. A sample label of the XML format is shown in Figure 30.

B. Files "TXT labels": the TXT version of the labels contain minimal details about the frames, a numeric representation of the label, an (x, y) location relative to the size of the frame and the height and width of the bounding box (relative to the frame size). A sample label of TXT format is shown in Figure 31.

ii. The remaining files are 361 extracted frames which contain anomalous events such as "Fight" and "Crowd Surge". The frames are labelled in the formats previously discussed.

5. Folder "Italy": this folder contains footage of the Italy incident where an audience crowd heard a loud bang and dispersed hectically in one direction. There are two captured views of the incident. The footage publicly available only captures when the crowd has started to quickly disperse, and hence prevented the extraction of training or "normal" data. However, the footage is usable on non-specific scene modelling methods that do not require the training scene to conform to the testing scene.

   (a) Folder "View_1": this folder contains the footage of the incident shot by a reporter from an eye-level view of the crowd. The footage is also extracted into frames for convenience. As previously mentioned, there is no training ("normal") data for this scene.

      i. Folder "Test": this folder contains the video format of the incident where the crowd disperses to the west side of the video and the extracted video frames.

         A. File "Footage.mp4": this video file is the captured footage of the entire scene.

         B. Files: the remaining files are the 702 extracted frames; all these frames are "abnormal".

   (b) Folder "View_2": this folder contains the footage of the incident captured by a CCTV camera from a high angled view of the crowd. The footage is also extracted into frames for convenience. As previously mentioned, there is no training ("normal") data for this scene.

      i. Folder "Test": this folder contains the video format of the incident where the crowd disperses to the east side of the video and the extracted video frames.

         A. File "Footage.mp4": this video file is the captured footage of the entire scene.

         B. Files: the remaining files are the 702 extracted frames; all these frames are "abnormal".

6. File "Read Me.txt": this file contains a summarised version of the aforementioned details of the dataset. The content of this file is included in the next section (Section 5.1.4) as well as more details and sample images for each scene.

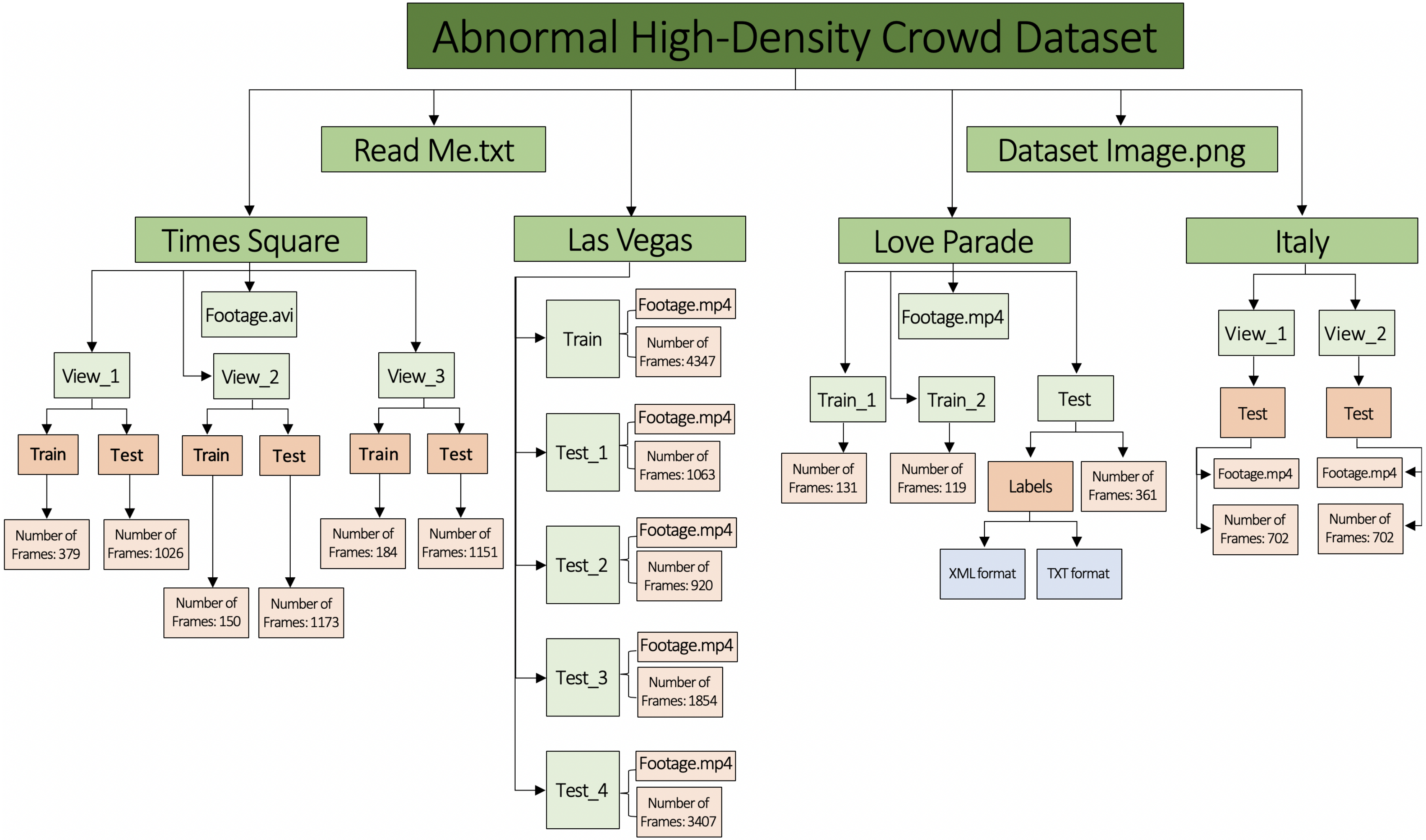


Figure 36: Detailed structure of the Abnormal High-Density Crowd - AHDCrowd dataset

5.2.1 Dataset View/Usage Statistics

The following statistics present the activity of the dataset after it was created and published on https://www.kaggle.com/ in December 2019. These statistics demonstrate the current activity state of the dataset as of May 2021:

- Number of views: 3515
- Number of downloads: 166

Moreover, contact has been established with a computer vision researcher who has utilised this dataset for object detection and tracking (pedestrians, cars, motorcycles, bicycles, etc.) within multiple views of the same environment. The research utilising this dataset has yet to be published.

5.3 Usage and Evaluation Protocols

There are a variety of applications in which this dataset can be utilised. The specific usage and evaluation methods applicable to this dataset are documented below.

5.3.1 Usage

The uses of this novel high-density crowd dataset are mainly for research in the computer vision field. The main use of this dataset is the detection of anomalous behaviour within a highly crowded environment. Details of several fields that can utilise this dataset in experimentation are noted below:

- **Crowd density estimation:** this dataset can be utilised to estimate the density of a crowd, more high-density datasets are required in the crowd density estimation field. The normal footage includes highly dense crowd walking around (at a normal pace), this footage can be utilised in crowd density estimation architectures. However, additional ground-truth data (estimated number of people in each frame of the crowd video) will need to be generated to be able to evaluate crowd counting and density estimation methods.
- **Tracking and re-identification:** as previously noted this dataset has been utilised in tracking and re-identification of specific objects. There is a gap in the availability of multi-view high-density crowd footage. This dataset adheres to both constraints, qualifying it to be utilised in the tracking and re-identification field. However, ground-truth data of the trajectories of specific objects (individuals, cars, etc.) needs to be generated to allow tracking and re-identification methods to be evaluated accurately.
- **Crowd anomaly detection:** as one of the main contributions of this research, the

gap regarding datasets with combined features of high-density crowds, annotations and occurrences of anomalous behaviour has been addressed. This dataset, at the time of writing this thesis, is the only high-density crowd dataset containing annotated anomalous behaviour. The field of crowd behaviour analysis and anomaly detection can utilise this dataset as a benchmark dataset for the evaluation of their methods in a high-density crowd environment.

5.3.2 Evaluation Protocols

The annotations documented for the anomalous occurrences in each video of this dataset (details of the annotation process in documented Section 5.1.3) enables researchers to evaluate their methods using various evaluation metrics such as Accuracy, Recall, Precision, F1 Score, Mean Square Error, ROC Curves, Equal Error Rate and Area under the ROC Curve (described in Section 3.5).

As a preliminary evaluation of this dataset, the amount of training/testing frames in each scene of the benchmark datasets in the crowd behaviour anomaly detection methods are compared to the amount of training/testing frames for each scene in the dataset produced by this research and the amounts are similar in range. Additionally, practical usage of this dataset is documented in Chapter 6 to demonstrate the dataset is used to train and test state-of-the-art low to medium-density crowd anomaly detection methods and evaluate their performance on high-density crowds.

5.4 Challenges and Limitations

The challenges and limitation experienced throughout the creation of this abnormal high-density crowd dataset are detailed below.

5.4.1 Challenges

Multiple challenges were faced when collecting footage to create the proposed dataset. Some of these challenges are:

- Scarcity of highly dense crowd footage.
- Due to the sensitive nature of anomalous incidents (fights, stampedes, etc.) footage of such incidents are not always publicly available.
- Captured footage is usually unstable (e.g. captured on a camera phone), generating videos that are inadequate for training a model for anomaly detection.
- The above challenges also contributed to the limited variety of anomalous behaviour types (panicked dispersion, fight and crowd surge) in the footage found.

5.4.2 Limitations

There are four major limitations in this dataset that could be addressed in future research:

- There are unlimited types of anomalous behaviour in a crowd. However, footage capturing these anomalies is either not public or non-existent. Limiting the diversification of anomaly types.
- Due to the previous limitation, anomalous behaviour in this dataset, as well as other benchmark datasets does not allow crowd behaviour anomaly detection methods to definitively claim their method can detect any/all anomalous behaviour presented to the model.
- The ground-truth data for crowd counting needs to be generated to allow the ideal testing of crowd counting and density estimation methods.
- Generation of the ground-truth data for tracking throughout different views is required. This will allow accurate testing by tracking and re-identification methods.

5.5 Conclusion

While reviewing methods for crowd analysis, the necessity for benchmark datasets became apparent. Fields such as crowd counting, density estimation, tracking, person re-identification and crowd anomaly detection all require benchmark datasets for their experimentation to achieve consistency. Benchmark datasets such as UMN, Avenue and UCSD are consistently used for anomaly detection within low to medium density crowds. Existing methods have not been analysed through application to a high-density crowd due to the lack of availability of an anomalous high-density crowd dataset until now. The AHDCrowd dataset produced in this research fills this gap. This dataset was produced by collecting, processing and labelling footage of environments containing highly dense crowds and occurrences of anomalies. The veracity of the annotation processes was validated by several researchers in the computing research field. The challenges and limitation of the dataset have also been noted to be addressed in future research. Preliminary evaluation of the dataset through a comparison of the amount of training/testing data against benchmark datasets suggests that the produced dataset (Abnormal High-Density Crowd) can be used to test state-of-the-art crowd anomaly detection methods, as well as the novel anomaly detection method proposed in this research. The evaluation of this is discussed in Chapter 6.

6 Experiments and Results

We carried out the following experiments to evaluate the contributions achieved in this research. The first experiment includes the application and evaluation of state-of-the-art crowd anomaly detection methods applied to the Abnormal High-Density crowd dataset - AHDCrowd (Mahmoud and Arafa, 2020). Standard evaluation methods have been applied to determine the performance of anomalous behaviour detection methods in a high-density environment. The second crucial experiment is the application of the proposed novel crowd anomaly detection method incorporating several motion representations such as Dynamic Images with conditional generative adversarial networks (GANs). Extensive experiments were applied on three benchmark datasets to validate the effectiveness and efficiency of the proposed method in comparison to the state-of-the-art in anomaly detection. The last experiment includes the performance evaluation of the proposed crowd anomaly detection method on several scenes from the Abnormal High-Density Crowd - AHDCrowd dataset.

6.1 Abnormal High-Density Crowd Dataset

In this section, details and results of applying several abnormal crowd behaviour detection methods to high-density crowd are documented. All experiments applied in this section were implemented on Google Colab (Mahmoud, 2020).

6.1.1 Crowd anomaly detection methods

Current methods for crowd anomaly detection architectures have been chosen to train and test using the AHDCrowd dataset. The methods documented below have all been trained and tested on low to medium density crowd footage by their authors. However, as a contribution of this research the AHDCrowd (Mahmoud and Arafa, 2020), containing occurrences of anomalous behaviour within high-density crowd footage, is used to evaluate the performance of the selected methods on high-density crowds. The methods being evaluated are Spatiotemporal Autoencoder (Chong and Tay, 2017), Future Frame Prediction (Liu et al., 2018b), and Anomaly Detection Using Multilevel Representations (Vu et al., 2019) the details are documented below.

6.1.1.1 Abnormal Event Detection in Videos using Spatiotemporal Autoencoder

Following the work presented by (Chong and Tay, 2017), detection of anomalies within a crowd is achieved using Spatiotemporal Autoencoders. To test the AHDCrowd dataset, the Spatiotemporal Autoencoders are applied using the settings and parameters provided in(Chong and Tay, 2017). These setting are applied to evaluate the performance of Spatiotemporal Autoencoders on high-density crowds as opposed to low and medium-density crowds. Initially, the input data (crowded scene) is pre-processed to be ready for the training stage. Pre-processing has three stages:

- Resize extracted frames to a resolution of 227 x 227 for consistency.
- Frame pixels are all scaled between 0-1.
- Extracted frames are converted to greyscale and normalised to have mean and unit variance values of zero.
- Extracted frames are split into temporal sequences of 10 frames using a sliding window method with several skip strides.
- The size of the training data is increased in the temporal dimension by applying data augmentation (Concatenating frames with stride-1, stride-2 and stride-3).

To build and train the convolutional long short term memory (LSTM) autoencoder network Keras is used, Figure 37 is an illustration of the architecture built. There are two parts to the network, a spatial auto-encoder and a temporal encoder/decoder. They are used to encode the spatial features of the input frames then it is fed as input to the temporal encoder/decoder to encode the temporal features extracted. The temporal encoder/decoder consists of a three-layer convolutional LSTM and the spatial encoder/decoder contain two convolution and deconvolution layers successively.

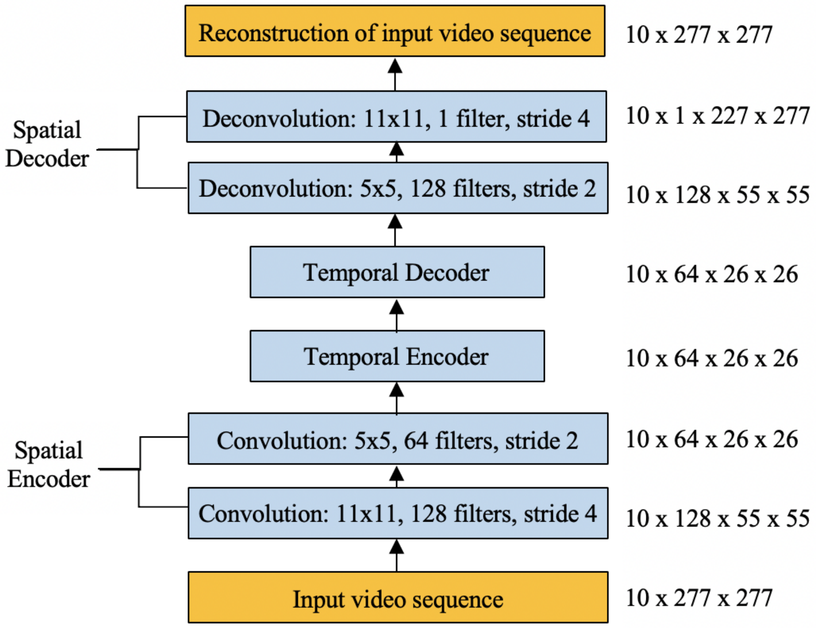


Figure 37: Stacked convolutional autoencoders with spatial and temporal encoder/decoder. Adapted from (Chong and Tay, 2017)

An Adam optimiser is utilised and the learning rate for the network is set to 0.0001. The batch size = 8 and the network was trained for 50 epochs. To test and evaluate the network, a regularity score was calculated according to the equations noted in (Chong and Tay, 2017). The reconstruction error of a pixel's intensity value **I** is calculated using L2 norm (square root of the sum of squared vector values) for its corresponding **x,y** location in frame **t** as shown below:

$$e(x,y,t) = ||I(x,y,t) - fw(I(x,y,t))||2 \tag{23}$$

fw is an annotation for the previously trained model. The reconstruction error value of the whole frame is calculated by summing the reconstruction error of each pixel (**e(x,y,t)**):

$$e(t) = \sum_{(x,y)} e(x,y,t) \tag{24}$$

Then the sequence reconstruction cost (annotated as **src(t)**) for 10 frames is calculated using:

$$src(t) = \sum_{t'=t}^{t+10} e(t') \tag{25}$$

Finally, the abnormality score, **sa(t)**, is scaled between 0-1 using Equation 26. This is followed by calculating the regularity score, **sr(t)**, by subtracting the abnormality score from 1 and it is calculated using Equation 27.

$$s_a(t) = \frac{src(t) - src(t)_{min}}{src(t)_{max}} \tag{26}$$

$$s_r(t) = 1 - s_a(t) \tag{27}$$

6.1.1.1.1 Evaluation/Results:

The Spatiotemporal Autoencoder model by Chong and Tay (2017) is used to evaluate the AHDCrowd dataset produced in this research. The model was trained and tested on four incidents. The first incident is modelled using the UCSD Ped-1 dataset (to demonstrate the efficacy of this model), and the remaining three incidents are modelled using three scenes from the AHDCrowd dataset. The details and results of each experiment are documented below:

1. **UCSD Ped-1:**

   To demonstrate the ability of this method to detect an anomalous event in a low to medium density crowd the model was trained on all the UCSD Ped-1 training sets and tested on the 32nd scene of the dataset. As illustrated in Figure 38, the regularity scores calculated on all testing frames are graphically plotted with the frame number plotted against the X-axis and the regularity score plotted against the Y-axis. The decline in regularity scores (Chong and Tay, 2017) indicates the occurrence of an anomaly, the red circles bring attention to the abnormalities detected by the model. These detected abnormalities are consistent with the ground-truth data where two anomalies occur between frames 1-52 and 65-115, also shown in Figure 39. Both frames in Figure 39 show different instances of bicycles entering the scene which is considered as an anomaly in this dataset. Based on the plotting of the regularity scores a normality threshold of 0.875 can be suggested for this specific dataset to indicate the occurrence of an anomaly.

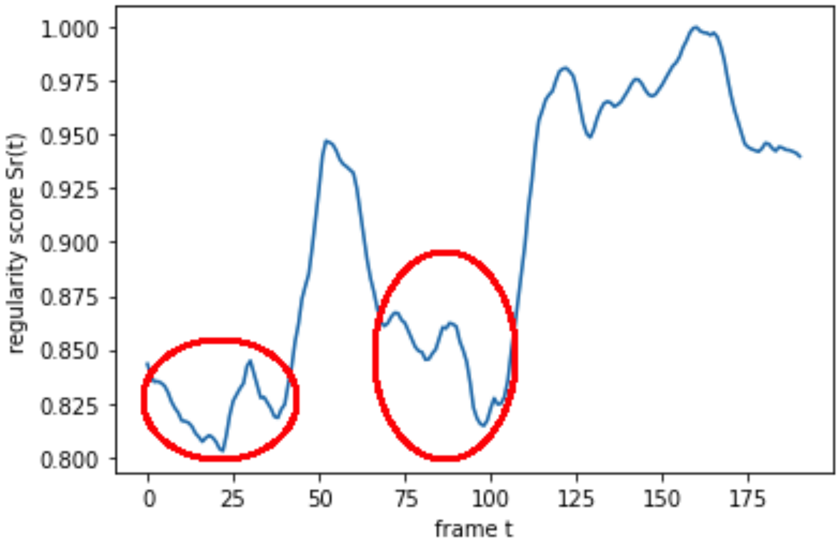


Figure 38: Regularity score (Sr(t)) and frame number (t) plotting results for the 32nd testing set in the UCSD Ped-1 dataset.

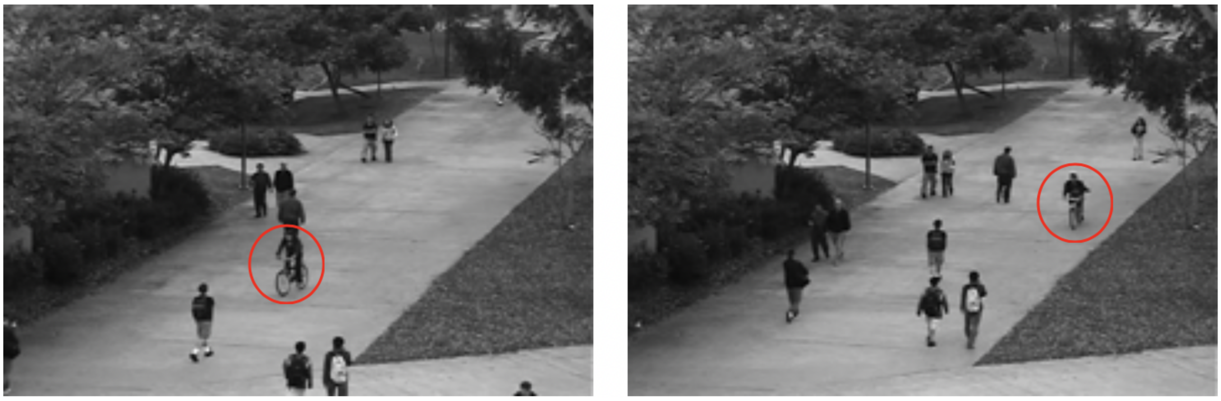


Figure 39: The sample images of the ground truth frames where a bicycle is driven through a walking path (anomaly) (Left: 27th frame, Right: 81st frame).

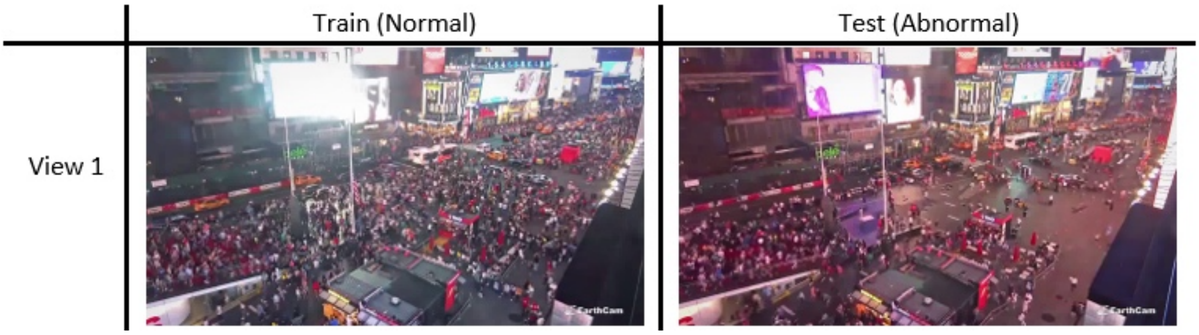


Figure 40: Sample images from each view angle of the Times Square incident (Mahmoud and Arafa, 2020).

2. **Abnormal High-Density Crowd (Times Square: View_1):**
   To analyse the ability of this method to detect anomalous events in a high-density crowd the model was trained on AHDCrowd. More specifically, the first view of the Times Square scene (sample frame shown in Figure 40), as illustrated in Figure 41 the regularity scores are plotted against the frame number. The ground truth data from the dataset account the start of an abnormality (people dispersing erratically) at frame 100 and the end at frame 1026. The chosen normality threshold based on when the anomaly begins (frame 100) is 0.900 (illustrated as a red line to divide normal and abnormal regularity scores in Figure 41). The results produced from the trained model suggests that there are occurrences where the frames return to a "Normal" state (higher than the specified threshold) illustrated as red circles. However, according to the ground truth data after the 100th frame, the crowd is in a continuous state of "Abnormal". As shown in Figure 42, an approximation of the ground truth regularity score plotting, the frames after the 100th frame should be plotted under the threshold line. The normality threshold chosen in this case is 0.919 corresponding to when the ground truth anomaly has started (frame 100). The results produced by this method when modelled on a highly dense crowd suggests

that the transition from abnormality detection with low to medium-density crowds into high-density crowds has weakened the performance. To further confirm if highly dense crowds decrease performance two more scenes have been utilised for training and testing.

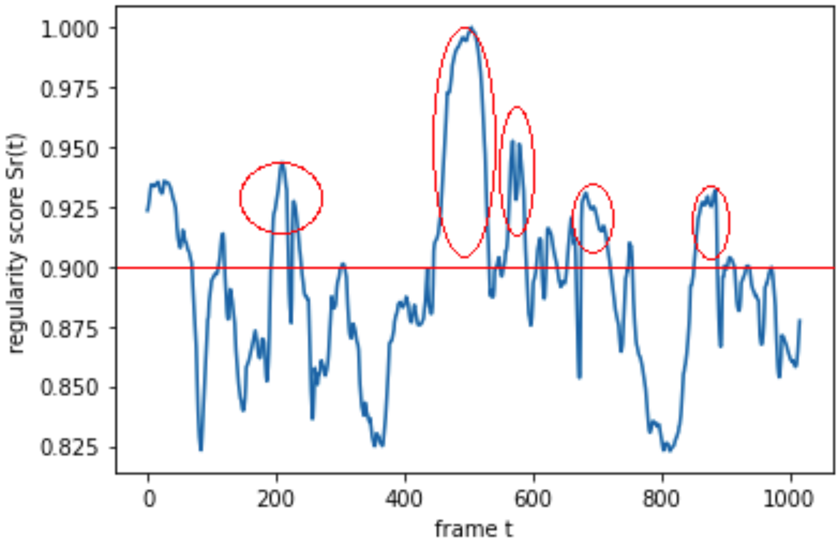


Figure 41: Regularity score plotting results modelled using Chong and Tay (2017) on the AHDCrowd (Times Square, View_1) dataset.

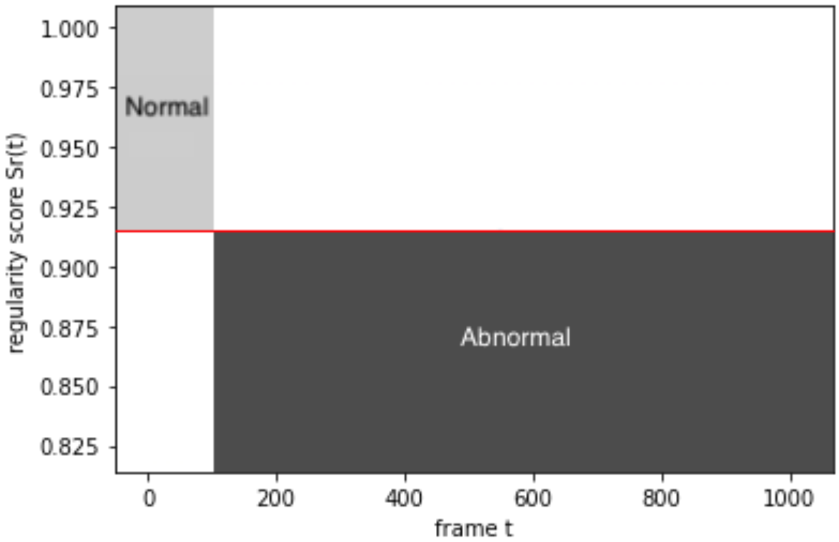


Figure 42: Estimated ground truth plotting using of the AHDCrowd (Times Square, View_1) dataset.

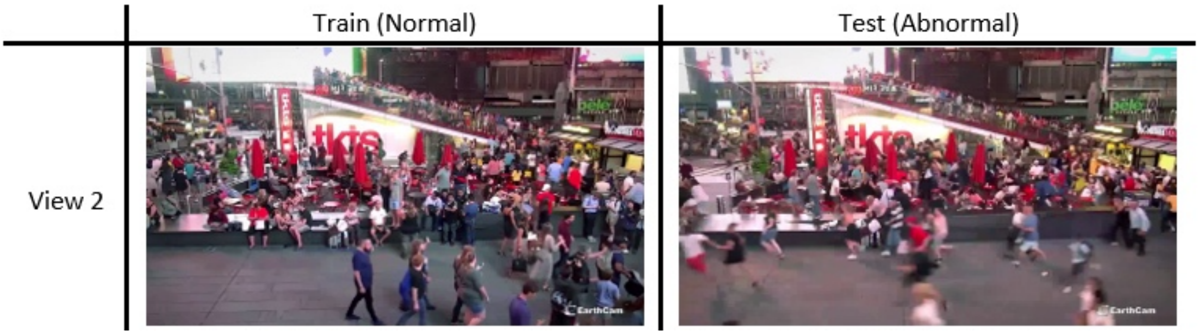


Figure 43: Sample images from each view angle of the Times Square incident (Mahmoud and Arafa, 2020).

3. **Abnormal High-Density Crowd (Times Square: View_2):**
   To continue the analysis of this method against high-density crowds another scene from the produced dataset was used to train/test model. In this experiment, the second view of the Times Square scene (sample frame shown in Figure 40) is used for modelling. The results produced by the method are plotted in Figure 44, the regularity scores are plotted against the frame number. The ground truth data from the dataset account the start of an abnormality (people dispersing erratically) at frame 30 and the end at frame 1173. The normality threshold, equivalent to 0.90, was determined based on the start of the anomalous behaviour (frame 30) for this scene. The threshold is illustrated as a red line to divide normal and abnormal regularity scores in Figure 44. The results generated from the trained model shows occurrences where the regularity scores exceed the specified threshold in the frame range 700 to 1173. This indicates a return to a "Normal" state, illustrated as red circles. However, according to the ground truth data, after the 30^{th} frame, the crowd is in a continuous "Abnormal" state. As shown in Figure 45, an estimated plotting of the ground truth regularity scores, the frames after the 30^{th} frame should be under the threshold line. The normality threshold in the case of the ground truth data plotting is 0.90, this is computed based on the ground truth of when an anomaly has begun (frame 30). The results produced by this method when modelled on this scene demonstrates better results than the previous scene. The previous experiment showed five peaks of normality that do not conform with the ground truth data in the frame range of 200 to 1000. Whereas this experiment shows two major instances of the regularity score increasing above the threshold line, in a frame range of 700 to 1200, that do not conform with the ground truth data. The next experiment utilises a scene where a fight takes place within a high-density crowd to evaluate if the model can detect localised anomalies on a frame-level basis.

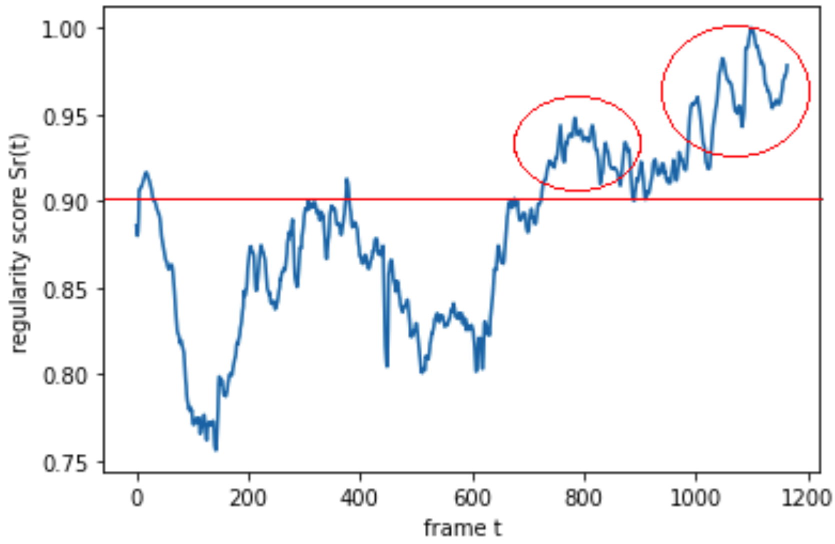


Figure 44: Regularity score plotting results modelled using using Chong and Tay (2017) on the AHDCrowd (Times Square, View_2) dataset.

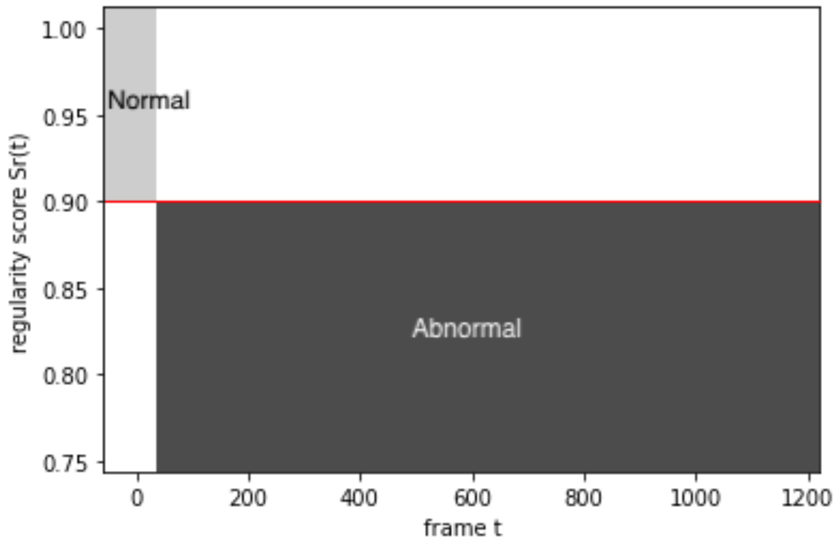


Figure 45: Estimated ground truth plotting of the AHDCrowd (Times Square, View_2) dataset.

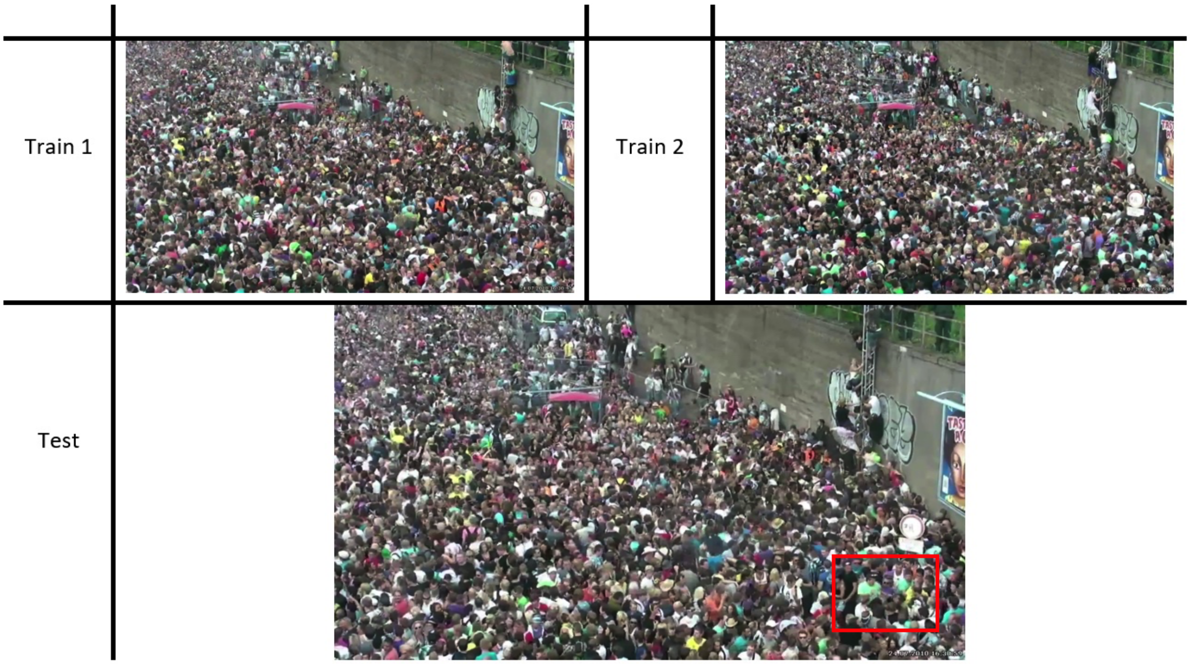


Figure 46: Sample images from the Love Parade incident, the anomaly is located using a red bounding box (Mahmoud and Arafa, 2020).

4. **Abnormal High-Density Crowd (Love Parade):**
The last experiment applied using this method was modelled on the Love Parade incident in the AHDCrowd dataset, the model was trained on Train_1 of the dataset (sample frame shown in Figure 46). The plotted regularity score results are illustrated in Figure 47, the regularity scores are plotted against the frame number. The ground truth data from the dataset account the start of an abnormality (a small group fighting) at frame 20 and the end at frame 300, the remaining frames are considered as "Normal" since the fight has ended. The normality threshold was determined to be 0.985 based on when the anomalous behaviour has started (frame 20) in this scene. The threshold is illustrated as a red line to divide normal and abnormal regularity scores in Figure 41. The generated results from the trained model indicate occurrences of anomalies where the regularity scores exceed the chosen threshold in the frame range 250 to 300. These occurrences, illustrated as red circles, do not conform with the ground truth data. The ground truth data indicates that the anomaly begins at the 20th frame and ends at the 300th frame, this is considered as the "Abnormal" state. As illustrated in Figure 48, an estimation of the ground truth regularity scores plotting, the frames after the 20th frame should be under the threshold line until the frame 300. The normality threshold, computed based on the ground truth of when an anomaly has begun, is still 0.985. Results generated from this method after being modelled on this scene demonstrated better results than both previous scenes. This experiment demonstrates an improvement where only one major

instance of the regularity score increasing above the threshold line and not conforming to the ground truth. The frame range of 250 to 300 do not comply with the ground truth data that the abnormality continues until the 300th frame.

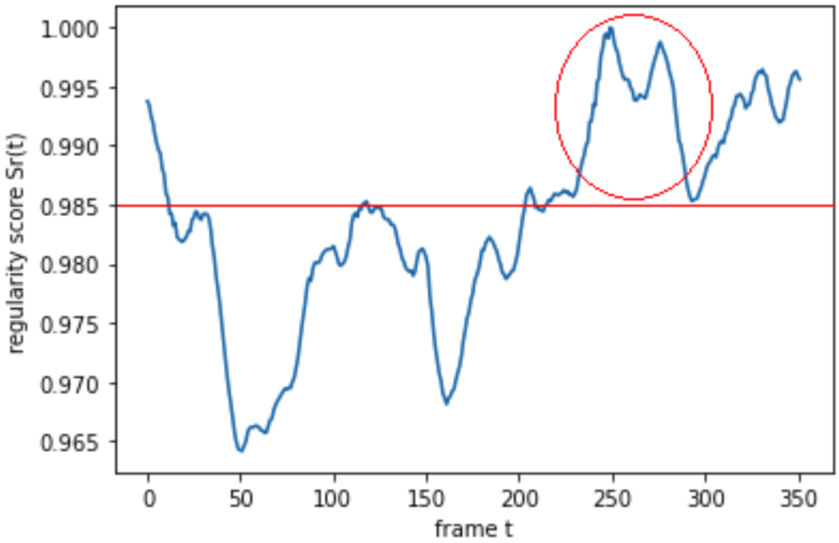


Figure 47: Regularity score plotting results modelled using using Chong and Tay (2017) on the AHDCrowd (Love Parade) dataset.

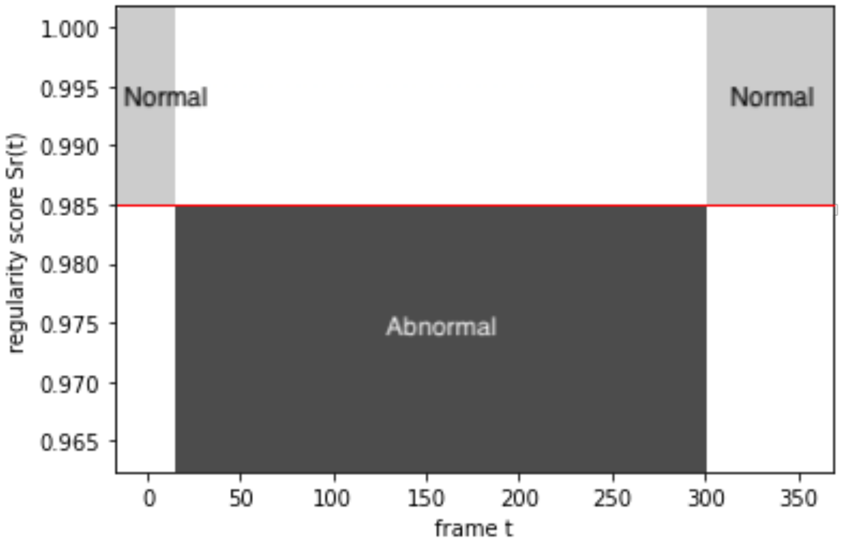


Figure 48: Estimated ground truth plotting of the AHDCrowd (Love Parade) dataset.

To obtain a better understanding of what the previously noted results mean in comparison to other methods, more experimentation has been applied. Another state-of-the-art method has been chosen to be trained and tested using the AHDCrowd dataset created in this research. The details of the method, the experiment and the generated results are discussed below.

6.1.1.2 Future Frame Prediction for Anomaly Detection – A New Baseline

Documented in Section 3.3.4 the research presented by Liu et al. (2018b) was reviewed as one of the state-of-the-art methods in crowd anomaly detection that utilised Generative Adversarial Networks (GANs) as a part of their architecture. For this experiment, the architecture illustrated in Figure 49 was modelled on multiple scenes from the AHDCrowd dataset, the details of the algorithm and produced results of this experiment are detailed below.

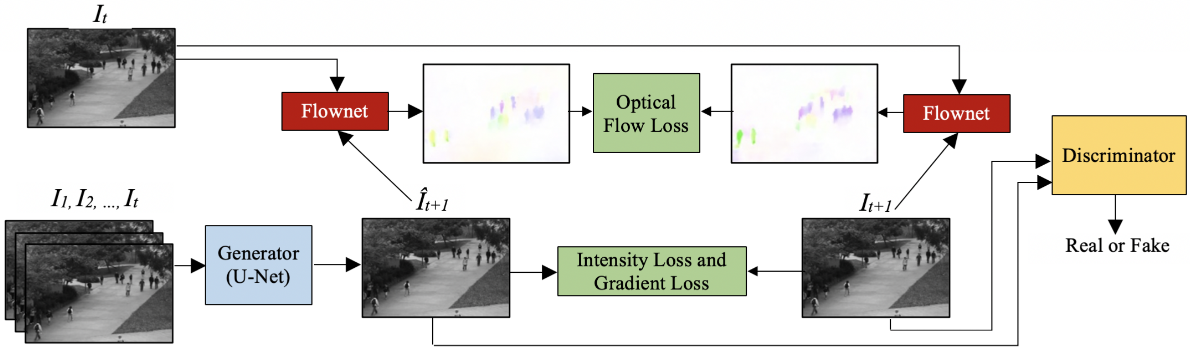


Figure 49: Future frame prediction for anomaly detection framework. Adapted from (Liu et al., 2018b)

The main idea behind this method, which utilises Generative Adversarial Networks are better (explained in Chapter 4), is to detect anomalous behaviour in an image frame of a crowd by trying to predict said frame, if the frame is equivalent to the ground truth then no anomalies are detected. Whereas, if the predicted frame is not equivalent (based on a predetermined threshold) then an anomaly is detected. Following the work by Liu et al. (2018b) the input frames are given to a U-Net generator (**G**) (based on (Isola et al., 2017)) illustrated in Figure 50. The input and output frames are the same resolution, and in comparison to autoencoders, this structure produces images that are clearer. **G** is trained to produce images that the discriminator **D** categorises as genuine, the adversarial training loss function used in training **G** is denoted in Equation 28. Mean Square Error loss function (L_{MSE}) is utilised and denoted below in Equation 29. Ground truth future frame prediction is denoted as I_{t+1} and the model future frame prediction is denoted as $\hat{I}_{t+1}$.

$$L^G_{adv}(\hat{I}) = \sum_{i,j} \frac{1}{2} L_{MSE}(D(\hat{I})_{i,j}, 1) \tag{28}$$

i, j represent the extracted spatial patches.

$$L_{MSE}(\hat{Y}, Y) = (\hat{Y} - Y)^2 \tag{29}$$

Y is given values in $\{0,1\}$ and $\hat{Y}$ in $\epsilon[0, 1]$.

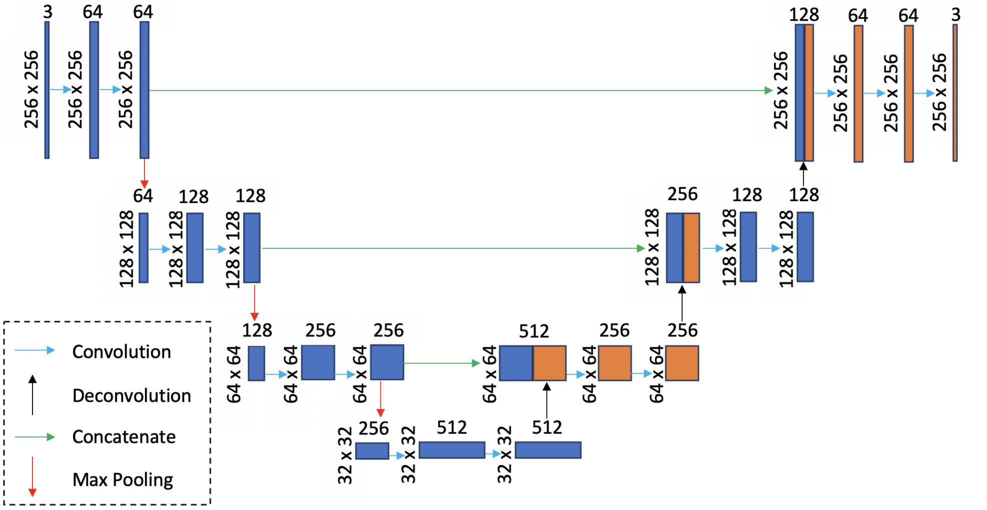


Figure 50: U-Net architecture network for prediction. Adapted from (Liu et al., 2018b)

On the other hand, the discriminator (**D**) is trained to discriminate if the frame given from **G** is fake or genuine. Based on (Isola et al., 2017) the discriminator is a patch discriminator; the scalar outputs from **D** are equivalent to an input frame patch. To compute the adversarial training MSE loss of the discriminator the following equation is used:

$$L^{D}_{adv}(\hat{I}, I) = \sum_{i,j} \frac{1}{2} L_{MSE}(D(I)_{i,j}, 1) + \sum_{i,j} \frac{1}{2} L_{MSE}(D(\hat{I})_{i,j}, 0) \tag{30}$$

The method utilises four constraints when training: intensity and gradient constraints, motion constraint and the adversarial training constraint (detailed above). The intensity constraint assures all frame pixels are similar in the RGB space by minimising the $L2$ distance between the predicted and ground truth frames, denoted as $\hat{I}$ and I respectively, in the intensity space. To compute the intensity loss the following equation is used:

$$L_{int}(\hat{I}, I) = \left|\left|\hat{I} - I\right|\right|_2^2 \tag{31}$$

The second constraint, gradient, is calculated using the gradient loss equation from:

$$L_{gd}(\hat{I}, I) = \sum_{i,j} \left|\left| |\hat{I}_{i,j} - \hat{I}_{i-1,j}| - |I_{i,j} - I_{i-1,j}| \right|\right|_1 + \left|\left| |\hat{I}_{i,j} - \hat{I}_{i,j-1}| - |I_{i,j} - I_{i,j-1}| \right|\right|_1 \tag{32}$$

i, j represent the spatial index in a given frame.

The last constraint utilised in this method is the motion constraint, it is calculated using the temporal loss equation (Equation 33). The temporal loss is calculated based on the difference between the calculated optical flow of a predicted frame and the ground truth optical flow. A pre-trained CNN, Flownet (f), is used to estimate the optical flow of two frames.

$$L_{op} = \left|\left| f(\hat{I}_{t+1}, I_t) - f(I_{t+1}, I_t) \right|\right|_1 \tag{33}$$

The generators (**G**) loss is calculated by combining all the constraints detailed above using:

$$L_G = \lambda_{int} L_{int}(\hat{I}_{t+1}, I_{t+1}) + \lambda_{gd} L_{gd}(\hat{I}_{t+1}, I_{t+1}) + \lambda_{op} L_{op} + \lambda_{adv} L^G_{adv}(\hat{I}_{t+1}) \tag{34}$$

And to train the discriminator **D**, the following equation is used:

$$L_D = L^D_{adv}(\hat{I}_{t+1}, I_{t+1}) \tag{35}$$

The AHDCrowd dataset is tested using the Future Frame Prediction method settings and parameters given in (Liu et al., 2018b). These setting and parameters are applied to evaluate the performance of Future Frame Prediction on high-density crowds as opposed to low and medium-density crowds. The network is trained on the following specifications:

- The pixels of the input frames are normalised to [-1, 1].
- Input frames are resized to 256 x 256.
- Initially, t is set to 4 and 5 random consecutive frames are used.
- Adam optimiser is utilised for parameter optimisation.
- Grayscale input videos use a learning rate of 0.0001 for the generator and 0.00001 for the discriminator.
- Coloured input videos use a learning rate of 0.0002 for the generator and 0.00002 for the discriminator.
- The standard values for some hyper-parameters are $\lambda_{int} = 1.0$, $\lambda_{gd} = 1.0$, $\lambda_{op} = 2.0$ and $\lambda_{adv} = 0.05$.

Finally, to detect anomalies in new data (testing) Peak Signal to Noise Ratio (PSNR) is calculated using:

$$PSNR(I, \hat{I}) = 10 \log_{10} \frac{[max_{\hat{I}}]^2}{\frac{1}{N}\sum_{i=0}^{N}(I_i - \hat{I}_i)^2} \tag{36}$$

A high PSNR value for an input frame suggests the frame is normal. A regularity score($S(t)$) for each frame can then be calculated by normalising the PSNR values of all the input frames between [0, 1]. A set threshold for $S(t)$ determines if an input frame is normal or not. $S(t)$ is computed using:

$$S(t) = \frac{PSNR(I_t, \hat{I}_t) - min_t PSNR(I_t, \hat{I}_t)}{max_t PSNR(I_t, \hat{I}_t) - min_t PSNR(I_t, \hat{I}_t)} \tag{37}$$

6.1.1.2.1 Evaluation/Results:

Documented below are the results produced by training and testing the previously detailed Future Frame Prediction method on four different scenes. The first scene used is from the UCSD dataset (to demonstrate the efficacy of this model) and the remaining scenes are all from the AHDCrowd dataset produced in this research.

1. **UCSD Ped-2:**
   Initially, this method was trained on all 16 videos from the UCSD Ped-2 dataset to show the method's ability in the detection of anomalous events in a low to medium density crowd. The method was trained and tested based on the previously detailed configurations. Testing was applied to the 12 test videos of the dataset, the quantitative evaluation metrics used to measure the performance of the method are Equal Error Rate (EER), Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) curve. The values of each are noted below and shown in Figure 51.

   - AUC = 0.9539455634972204
   - EER = 0.11975308641975309
   - ROC:

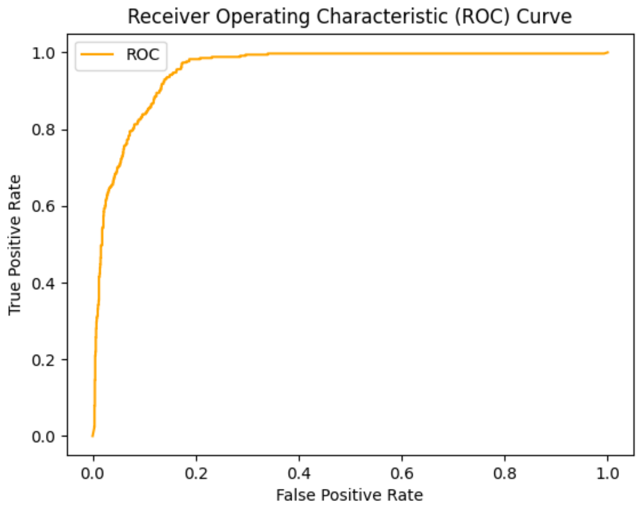


Figure 51: Receiver Operating Characteristic (ROC) curve plotting on UCSD Ped-2 dataset.

2. **Abnormal High-Density Crowd (Times Square: View_1):**
   To determine the abnormality detection performance of this method on a high-density crowd, the AHDCrowd dataset was used. The training configuration of the model on the Times Square: View_1 scene were 500 iterations on a batch size of 8. At the end of training the discriminator model had a global loss = 0.244921 and the generator model had a global loss = 0.092116766, an intensity loss = 0.0050, a gradient loss = 0.0706, an adversarial loss = 0.0061, a Flownet loss = 0.0105 and a PSNR error = 29.164692. The results noted below show a fairly acceptable AUC, EER and ROC (Figure 52), however in comparison to the achieved results on a low to medium density dataset, UCSD Ped-2 (tested above), these results show a significant decline in the performance of the method. To further analyse the anomaly detection capabilities of this method in a highly dense crowd, two more scenes from the AHDCrowd dataset were tested.

   - AUC = 0.8564948453608248
   - EER = 0.19567567567567568
   - ROC:

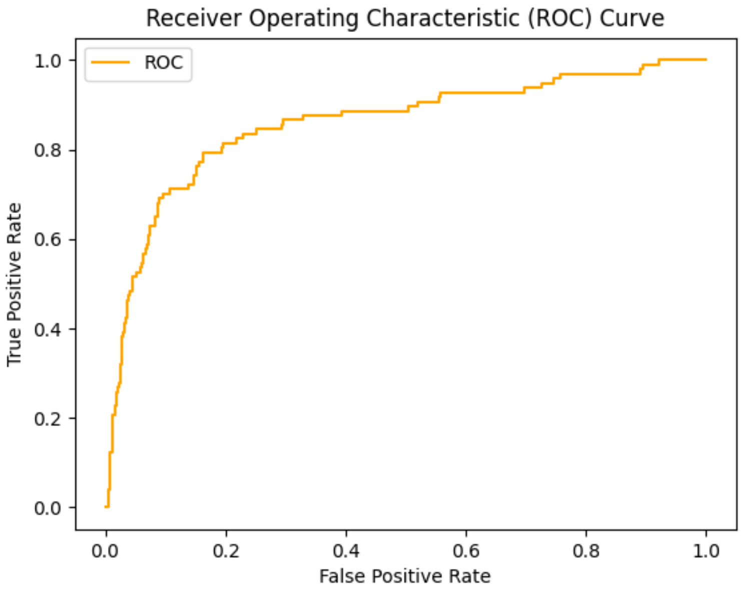


Figure 52: Receiver Operating Characteristic (ROC) curve plotting on AHDCrowd (Times Square: View_1) dataset using Liu et al. (2018b).

3. **Abnormal High-Density Crowd (Times Square: View_2):**

   To further analyse the performance of the future frame prediction method in the detection of anomalies within a high-density crowd this experiment utilises the AHDCrowd (Times Square: View_2) scene for training and testing. Unlike the previous training configurations, this experiment was applied for 800 iterations on a batch size of 4. At the 800th iteration the discriminator model had a global loss = 0.213080 and the generator model had a global loss = 0.1814959, an intensity loss = 0.0103, a gradient loss = 0.1017, an adversarial loss = 0.0080, a Flownet loss = 0.0615 and a PSNR error = 25.90276. The AUC EER and ROC (Figure 53) results of this experimentation show a significant improvement, to the previously documented results. The last experiment is applied on a scene where a fight takes place within a high-density crowd, this will determine if this method is able to detect localised anomalies on a frame-level basis.

   - AUC = 0.9856989030217376
   - EER = 0.0542432195975503
   - ROC:

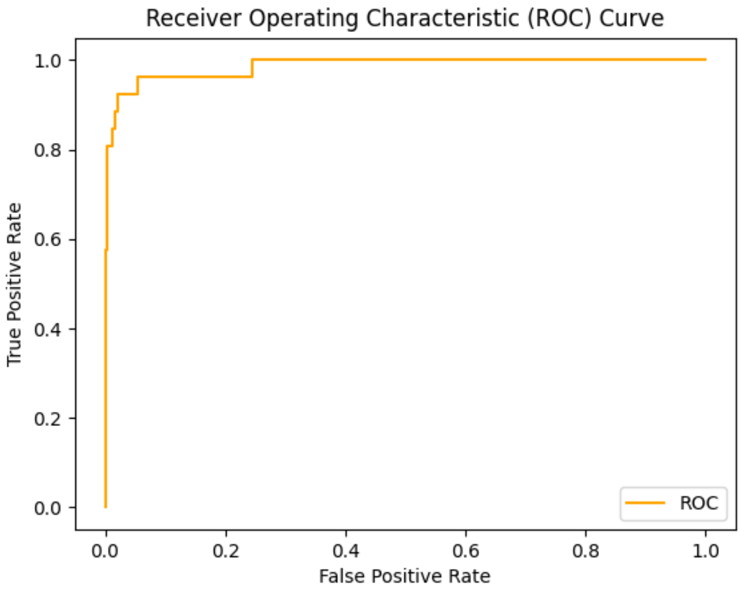


Figure 53: Receiver Operating Characteristic (ROC) curve plotting on AHDCrowd(Times Square: View_2) dataset using Liu et al. (2018b).

4. **Abnormal High-Density Crowd (Love Parade):**
   The final experiment is applied to test the ability of this method in detecting localised anomalous behaviour within a highly dense crowd. In this case, the anomaly is a fight and the training configuration are the same as the last experiment (800 iterations with batch size = 4). On the 800^{th} iteration the discriminator model had a global loss or 0.246807and the generator model had a global loss = 0.29538602, an intensity loss = 0.0272, a gradient loss = 0.1825, an adversarial loss = 0.0062, a Flownet loss = 0.0795 and a PSNR error of 22.942118. As demonstrated below and in Figure 54, the results on this scene have significantly decreased in comparison to all of the previous experiments. This is very likely due to the fact that the optical flow constraint of the method was not able to detect a major difference between consecutive frames. The set weight of the motion constraint (optical flow difference) has a significant impact on the outcome of the detection.

   - AUC = 0.5524118738404452
   - EER = 0.4714285714285714
   - ROC:

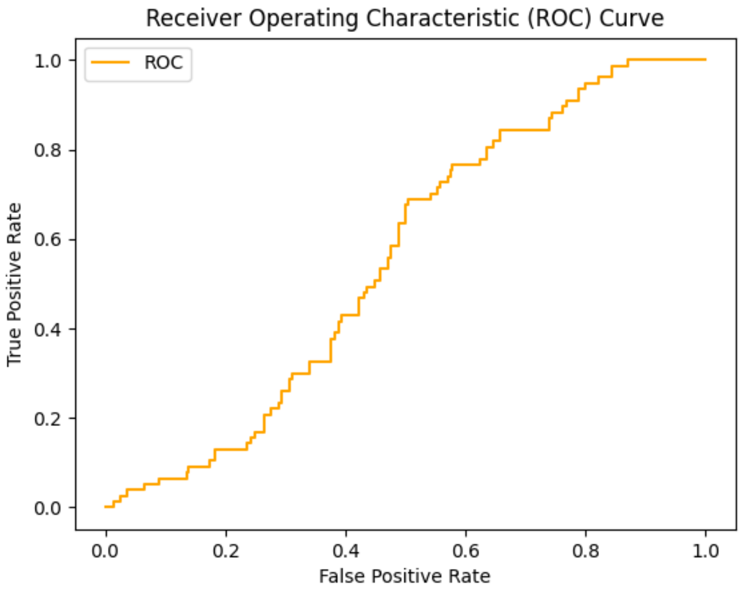


Figure 54: Receiver Operating Characteristic (ROC) curve plotting on AHDCrowd(Love Parade) dataset using Liu et al. (2018b).

6.1.1.3 Robust Anomaly Detection in Videos Using Multilevel Representations

Following the Anomaly Detection Using Multilevel Representations method by Vu et al. (2019), anomalies within crowds are detected using multilevel representations as previously reviewed in Section 3.3.4 and further detailed in Section 4.3. In this experiment, the method is trained and tested on multiple scenes from the AHDCrowd dataset as well as the UCSD Ped-2 benchmark dataset. Details of the method, experimental setup and results are documented below.

The architecture of this method is divided into two phases; training and detecting. The training phase, following (Vu et al. 2019) is applied by following these steps:

1. Input videos or data frames, $D_F = \{F_i\}_{i=1}^{N_f}$, with N_f as the extracted frames, the frames resized into 256 x 256 and scaled between [-1, 1].
2. Optical flow difference, O_i, is calculated for every two consecutive frames $(F_i,\ F_{i+1})$. Optical flow is originally computed using Brox et al. (2004), but in this research is computed using Sun et al. (2017) (further explained in Section 6.2).
3. DAE$_F$ and DAE$_O$ are denoising autoencoders trained on D_F and $D_O = \{O_i\}$ respectively, this is achieved by minimising the DAE Equation 19 (detailed in Section 4.3).
4. Encoding is applied by utilising convolutional layers with stride = 2 and kernel size = 5 x 5 then batch normalisation layers and leaky ReLU activation functions.

5. Decoding contains the same components as the encoding path but the convolutional layers are changed to deconvolutional layers.

6. Adagrad optimiser is used and $\gamma = 1$, the learning rate = 0.1. The network is trained for 500 epochs.

7. After DAE$_F$ is trained every frame F$_i$ is given to the network to achieve activations at every encoding layer.

8. To compute $F_i^{(l)}$ (l is the abstract representation level of the frame data), the activations are normalised to zero-mean and unit variance and clipped to [-1,1].

9. The previous step is applied again to compute $O_i^{(l)}$.

10. $D_F^{(l)} = \left\{F_i^{(l)}\right\}$ and $D_O^{(l)} = \left\{O_i^{(l)}\right\}$ are used to train the CGANs on the l^{th} level.

Two conditional GANs (CGANs) are trained on every level of representations following the steps by (Vu et al., 2019) and (Isola et al., 2017):

1. The CGAN $G^l_{F\to O}$ is used to generate the motion $O_i^{(l)}$ from the frame $F_i^{(l)}$ while the CGAN $G^l_{O\to F}$ is used to generate the frame from motion.

2. The network is set on a learning rate = 0.0002, $\lambda = 100$ and batch size = 1.

Similar to the previous experiments, the Robust Anomaly Detection method is tested using the AHDCrowd dataset. The same testing settings and parameters as (Vu et al., 2019) are used to evaluate the performance of the Robust Anomaly Detection method on high-density crowds as opposed to low and medium-density crowds. The testing or detection phase is based on single-level detection following these specifications:

1. The input frames, F_i, is used to compute the motion maps O_i the DAE$_F$ and DAE$_O$ utilise F_i and O_i to extract the high-level features $F_i^{(l)}$ and $O_i^{(l)}$.

2. The trained CGANs are given the high-level feature on every representation level to generate the motion and frame images $\hat{O}_i^{(l)} = G^{(l)}_{F\to O}\left(F_i^{(l)}, z\right)$ and $\hat{F}_i^{(l)} = G^{(l)}_{O\to F}\left(O_i^{(l)}, z\right)$.

3. $F_i^{(l)}$, $O_i^{(l)}$, $\hat{F}_i^{(l)}$ and $\hat{O}_i^{(l)}$ are set to zero in optical flow locations with a value of zero.

4. Generation maps are the calculated as $e^{(l)}_{F,i} = F_i^{(l)} - \hat{F}_i^{(l)}$ and $e^{(l)}_{O,i} = O_i^{(l)} - \hat{O}_i^{(l)}$ following the calculation described in Section 4.3.

5. The total error maps, $E^{(l)} = \left\{\bar{e}_i^{(l)}\right\}$, is then smoothed by averaging consecutive frames on a sliding frame window = 5.

6. A detection is made based on a set threshold β, when $\bar{e}_i^{(l)}(x,y)$ is bigger than β the binary detection map $D_i^{(l)}(x,y) = 1$ to indicate and anomaly and $D_i^{(l)}(x,y) = 0$ to

indicate normalcy.

6.1.1.3.1 Evaluation/Results:

Similar to the previous experiments this method was tested on four scenes; the benchmark dataset UCSD Ped-2 and three scenes from the Abnormal High-Density dataset. The scenes are Times Square: View_1, Times Square: View_2 and Love Parade. Training, using Vu et al. (2019), is based on the 32-16-8 network structure and applied for 500 iterations on each scene from the dataset. β is set to 0.8 and the method detects anomalies based on four different configurations; using features at all levels, using low-level features and top-level features, using only top-level features and using only low-level features. The configuration producing the best detection results are noted for each scene from the dataset. The generated frame-level detection results of training and testing this method on the various scenes are as follows.

1. **UCSD Ped-2:**
   Initially, this method was trained on 100 frames from a training video (normal footage) from the UCSD Ped-2 dataset to define normalcy. Additionally, this method is tested on the testing video (abnormal footage) from the UCSD Ped-2 dataset of the same scene to show the methods ability to detect anomalies in a low to medium density crowd. The method was trained and tested based on the previously detailed configurations and testing was applied to one of the test videos of the dataset. The quantitative evaluation metrics used to measure the performance of the method are Equal Error Rate (EER), Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) curve. The best detection results noted were achieved using *low-level features only*, noted below and in Figure 55:

   - AUC = 0.973125
   - EER = 0.030000
   - ROC:

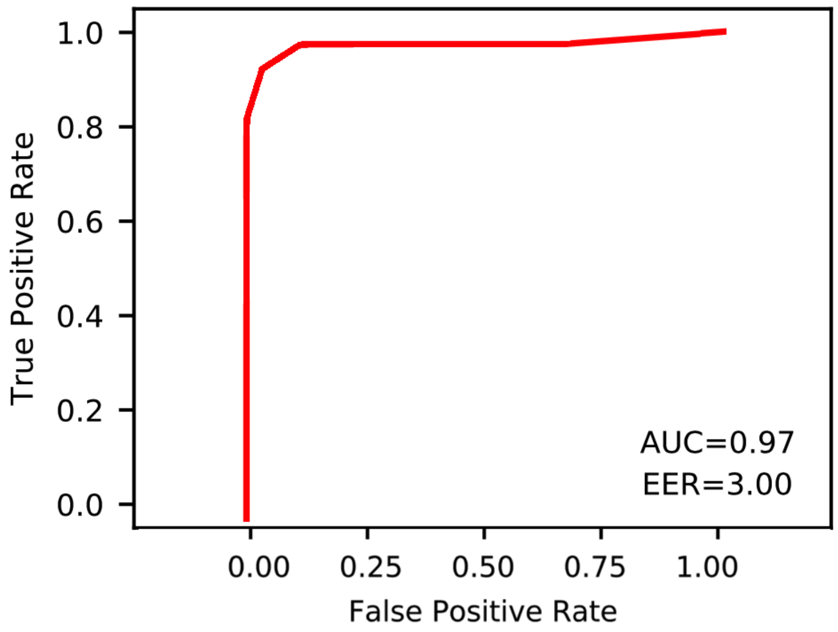


Figure 55: Receiver Operating Characteristic (ROC) curve plotting on UCSD Ped-2 dataset using Vu et al. (2019).

2. **Abnormal High-Density Crowd (Times Square: View_1):**
   To establish the performance of this method in the detection of abnormalities in a high-density crowd, the AHDCrowd dataset Times Square: View_1 was used. In training, 200 frames from the training segment of the dataset were used whereas in testing 150 frames were used. The 150 testing frames start with 50 frames of normal crowd behaviour and the remaining 100 frames contain anomalous behaviour. In comparison to the achieved results on a low to medium density dataset, UCSD Ped-2 (tested above), the testing results on this dataset presents a performance decline. The best-achieved detection results were produced using *features at top-level*, and AUC, EER and ROC (Figure 56) are noted below.

   - AUC = 0.574200
   - EER = 0.286667
   - ROC:

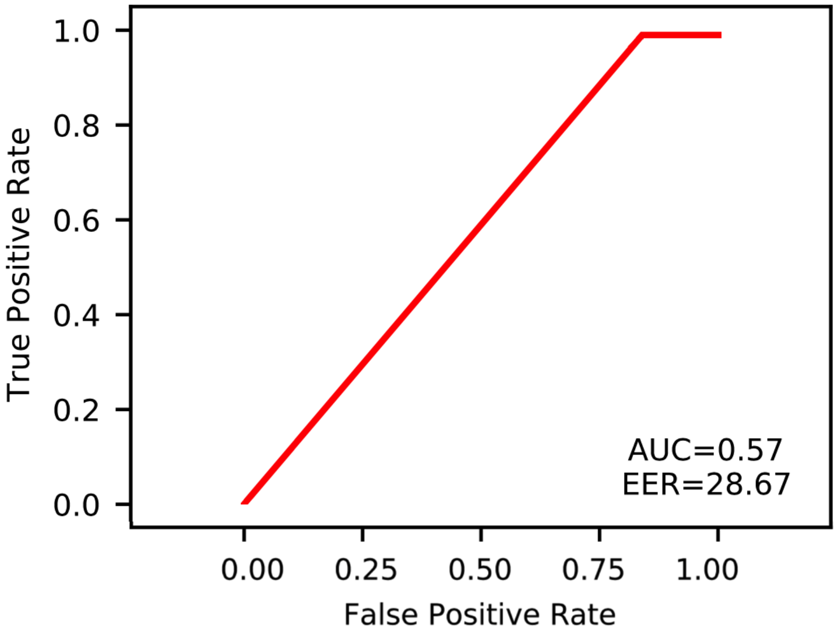


Figure 56: Receiver Operating Characteristic (ROC) curve plotting on AHDCrowd(Times Square: View_1) dataset using Vu et al. (2019).

To continue the analysis of the capabilities of this method to detect anomalies in a high-density environment two more scenes from the AHDCrowd dataset were used:

3. **Abnormal High-Density Crowd (Times Square: View_2):**
   As a continuation of the performance analysis of this method in a high-density crowd, this experiment utilises the AHDCrowd (Times Square: View_2) scene for training and testing. Unlike the previous experiment, 100 frames from the training and testing segments of the dataset were used and the abnormalities are present in the testing frames 30 to 100. The best detection results were produced using *features at low-level*, AUC, EER and ROC (Figure 57) results of this experimentation display significant higher AUC and EER results in comparison to the previous experiment. The last experiment is applied on a scene where a fight takes place within a high-density crowd, this will determine if this method is able to detect localised anomalies on a frame-level basis.

   - AUC = 0.660281
   - EER = 0.345455
   - ROC:

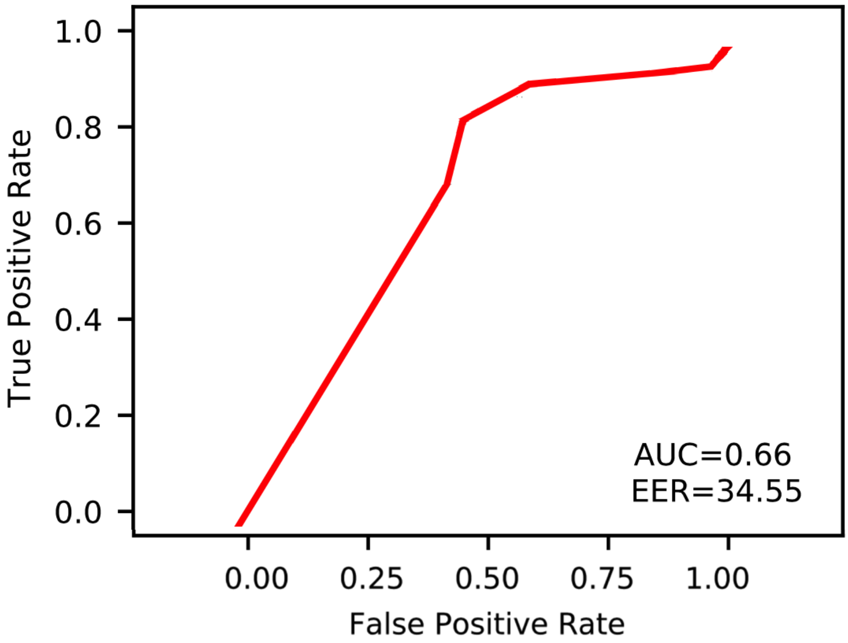


Figure 57: Receiver Operating Characteristic (ROC) curve plotting on AHDCrowd (Times Square: View_2) dataset using Vu et al. (2019).

4. **Abnormal High-Density Crowd (Love Parade):**

   The concluding experiment is applied to test localised anomaly detection capabilities of this method within a high-density environment. The localised anomaly in this instance is a fight within the crowd, and the training configurations are the same as the last experiment with 100 frames for both training and testing. The fight is shown in the testing frames 20 to 100 and the best results are produced using *features at top-level*. As noted below, the AUC, EER and ROC (Figure 58) experimental results on this dataset have significantly increased in comparison to all of the previous experiments.

   - AUC = 0.880856
   - EER = 0.163636
   - ROC:

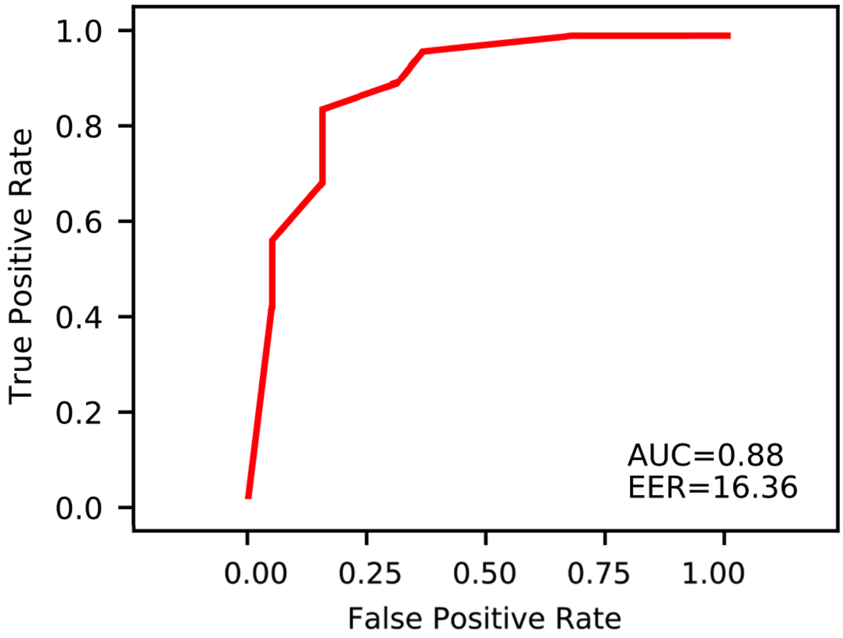


Figure 58: Receiver Operating Characteristic (ROC) curve plotting on AHDCrowd (Love Parade) dataset using Vu et al. (2019).

With the exception of the results produced from the (Liu et al., 2018b) method on the AHDCrowd (Times Square: View_2), the testing the methods by (Chong and Tay, 2017; Liu et al., 2018b; Vu et al., 2019) on this dataset produced performance results that are significantly lower than the results modelled on a low to medium crowd dataset (shown in Table 5). This demonstrates the limitations of these methods in transitioning from low to medium-density crowd anomaly detection into high-density crowd anomaly detection.

6.2 Optical Flow

Two optical flow estimation methods were applied in this research; Brox (Brox et al., 2004) and FlowNet (Sun et al., 2017) optical flow. The former method is the standard approach utilised in anomaly detection methods incorporating CGANs. The latter is a novel approach to calculate optical flow difference and is utilised as a substitute for the Brox method for the purpose of evaluating its effect on the performance of anomaly detection. Both methods have been applied on multiple crowd anomaly datasets; low to medium density crowds and high-density crowds. Finally, dynamic scenes are then evaluated. The qualitative results produced from each method is shown in each section and combined for easier viewing in Figures 67 and 68 when applied on benchmark and the AHDCrowd datasets respectively.

6.2.1 Brox Optical Flow

The Brox optical flow method (Brox et al., 2004) is used to compute the temporal development between two consecutive frames as previously discussed in Section 3.4.1. This method was applied to sample frames from the benchmark datasets UCSD Ped-1, UCSD Ped-2 and Avenue to illustrate the generated optical flow difference. Additionally, this method was applied to three scenes from the AHDCrowd dataset (Mahmoud and Arafa, 2020). Illustrations of the optical flow difference computed for the scenes in this dataset are shown in Figure 59 and Figure 60.

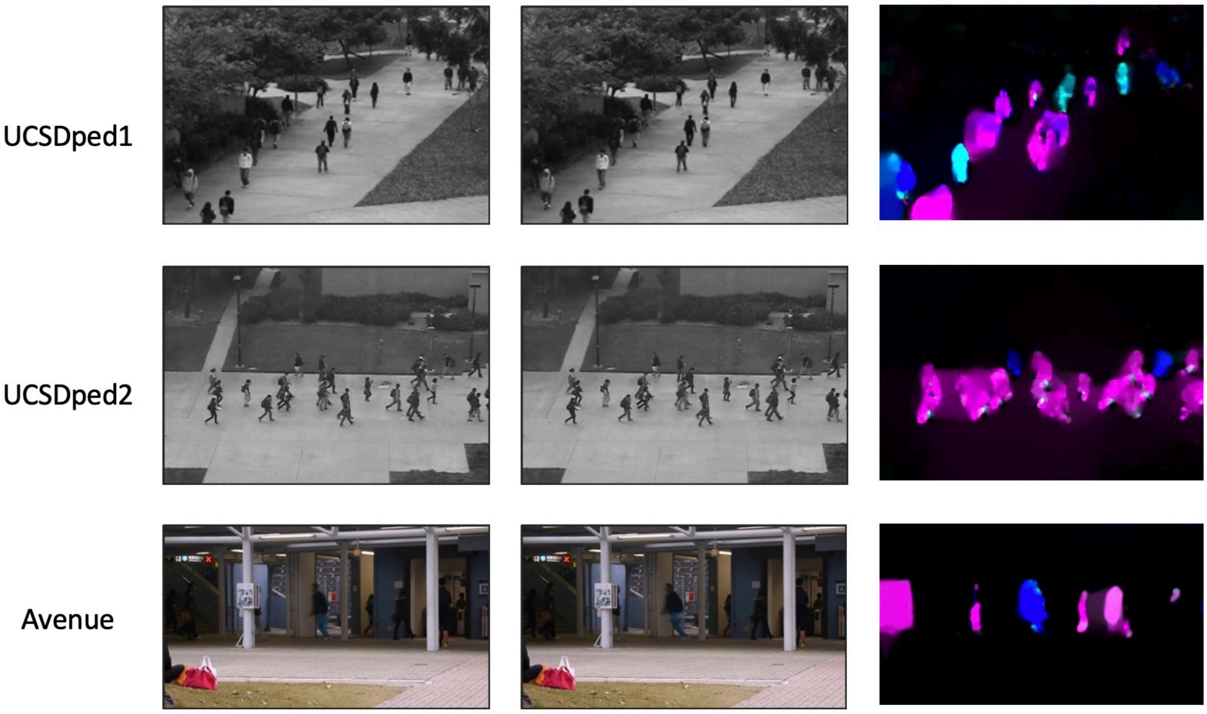


Figure 59: Three sample images from different benchmark datasets: each sample consists of two consecutive frames (first and second column). The third column is the result of the optical flow difference.

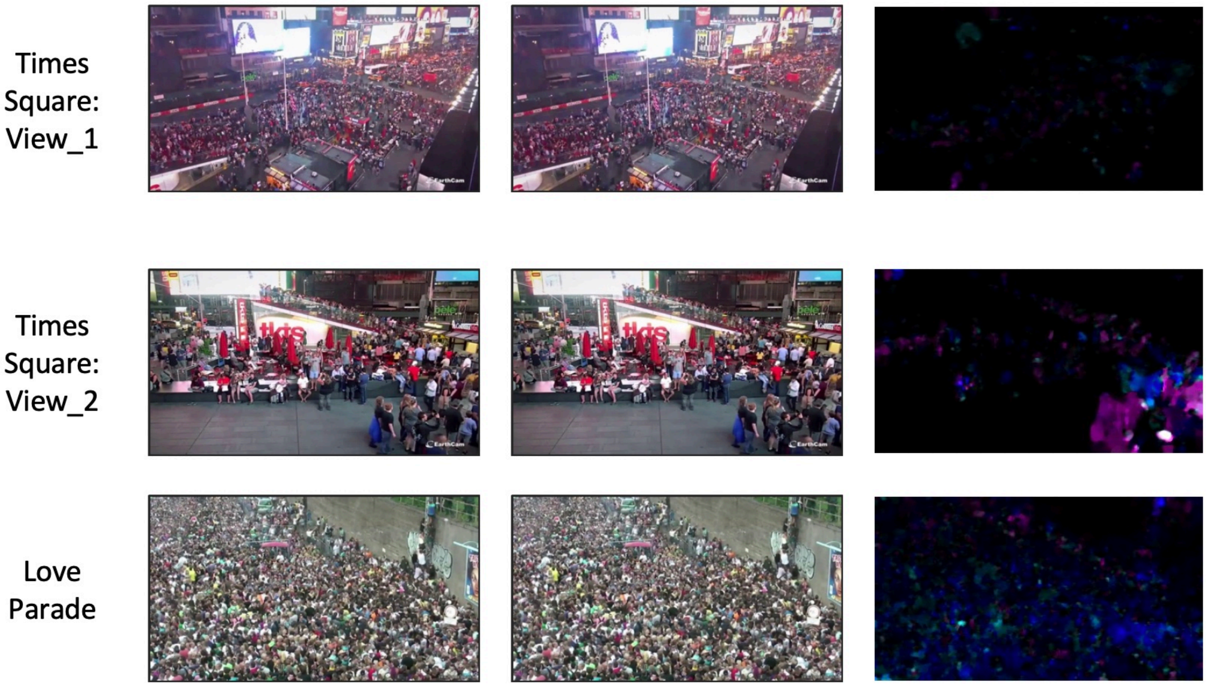


Figure 60: Three sample images from different scenes in the AHDCrowd dataset. Each sample consists of two consecutive frames (first and second column) and the third column is the result of the optical flow difference.

6.2.2 FlowNet Optical Flow

A more novel approach, FlowNet (Sun et al., 2017), for the calculation of optical flow difference was investigated. FlowNet is used to evaluate the performance difference between itself and Brox optical flow in conjunction with the proposed anomaly detection framework.

FlowNet (Sun et al., 2017), described in Section 3.4.1, generates optical flow estimation results with high accuracy and low running time. However, the method has difficulty predicting the optical flow difference between two consecutive frames when the magnitude between the two frames is large. The magnitude grows larger when objects are suddenly much farther than expected or when objects unexpectedly appear or vanish from frames. This method was applied to sample frames from the benchmark datasets UCSD Ped-1, UCSD Ped-2 and Avenue to illustrate the generated optical flow difference. The method was also applied to three scenes from the AHDCrowd dataset (Mahmoud and Arafa, 2020), the results for are shown in Figure 61 and Figure 62.

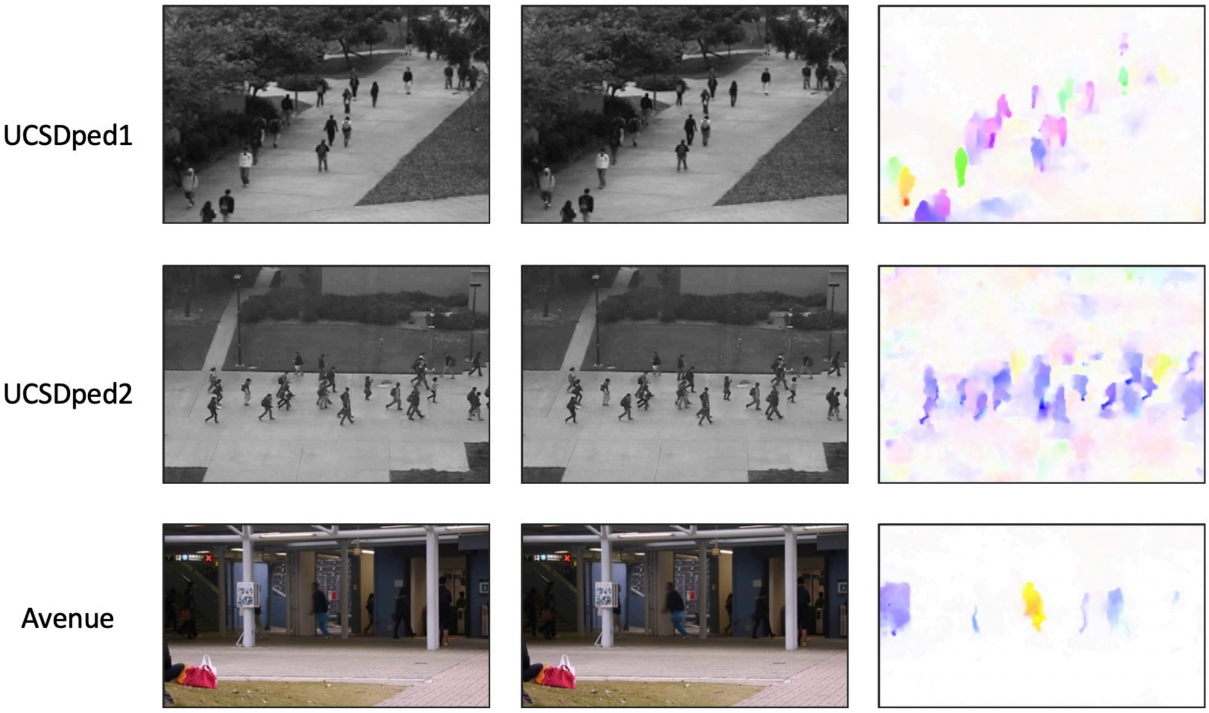


Figure 61: Three sample images from different benchmark datasets: each sample consists of two consecutive frames (first and second column). The third column is the result of the optical flow difference.

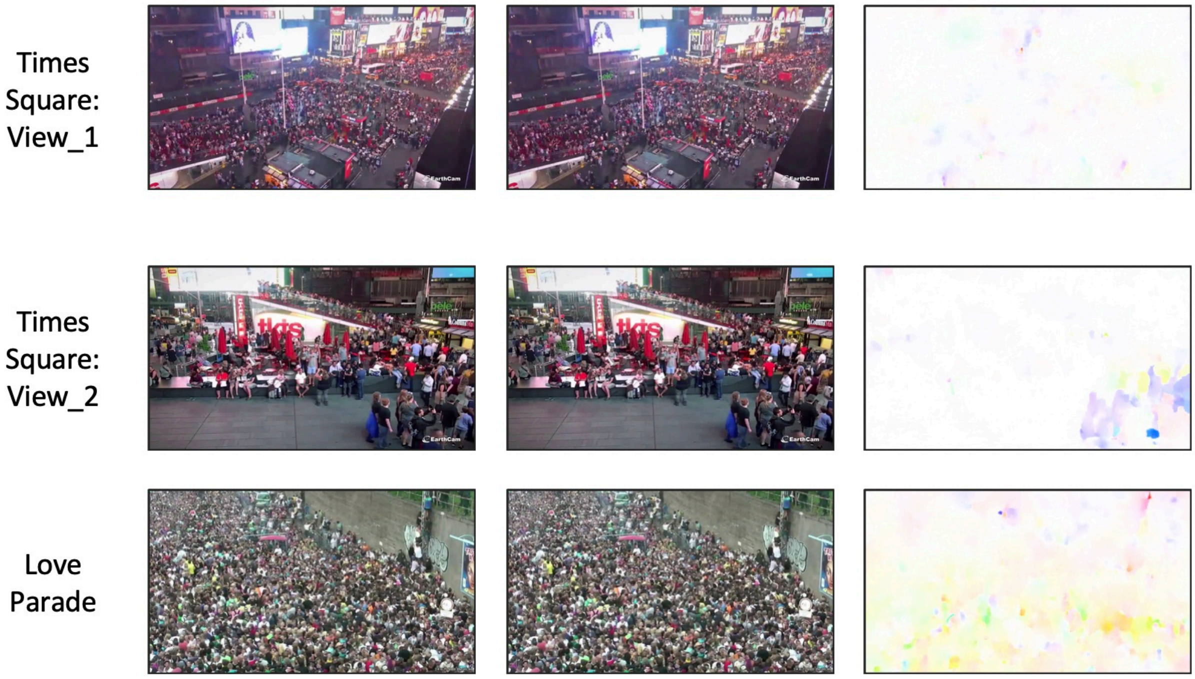


Figure 62: Three sample images from different scenes in the AHDCrowd dataset. Each sample consists of two consecutive frames (first and second column) and the third column is the result of the optical flow difference.

A limitation to using optical flow estimation methods is the methods can only extract the temporal development features between two consecutive frames for a set of frames. Better performance results can be achieved if temporal information over time (more than two frames) is extracted. Dynamic Images (Bilen et al., 2016) applies this theory. Better performance results have been substantiated in the field of action recognition, as previously noted in Section 3.4, using Dynamic Images. Implementation of Dynamic Images on different crowd anomaly detection methods is documented below.

6.2.3 Dynamic Images

The framework proposed in this research incorporates Dynamic Images (Bilen et al., 2016)) instead of the standard optical flow difference. While optical flow difference estimates the temporal difference between two consecutive frames, dynamic images (previously discussed in Section 3.4.2) incorporate the temporal changes throughout a set of consecutive images of size t represented as one image. The method was applied to sample frames from the benchmark datasets UCSD Ped-1, UCSD Ped-2 and Avenue to illustrate the dynamic image representation output with $t = 50$. The method was also applied to three scenes from the AHDCrowd dataset (Mahmoud and Arafa, 2020), the results for are shown in Figures 63 and 64.

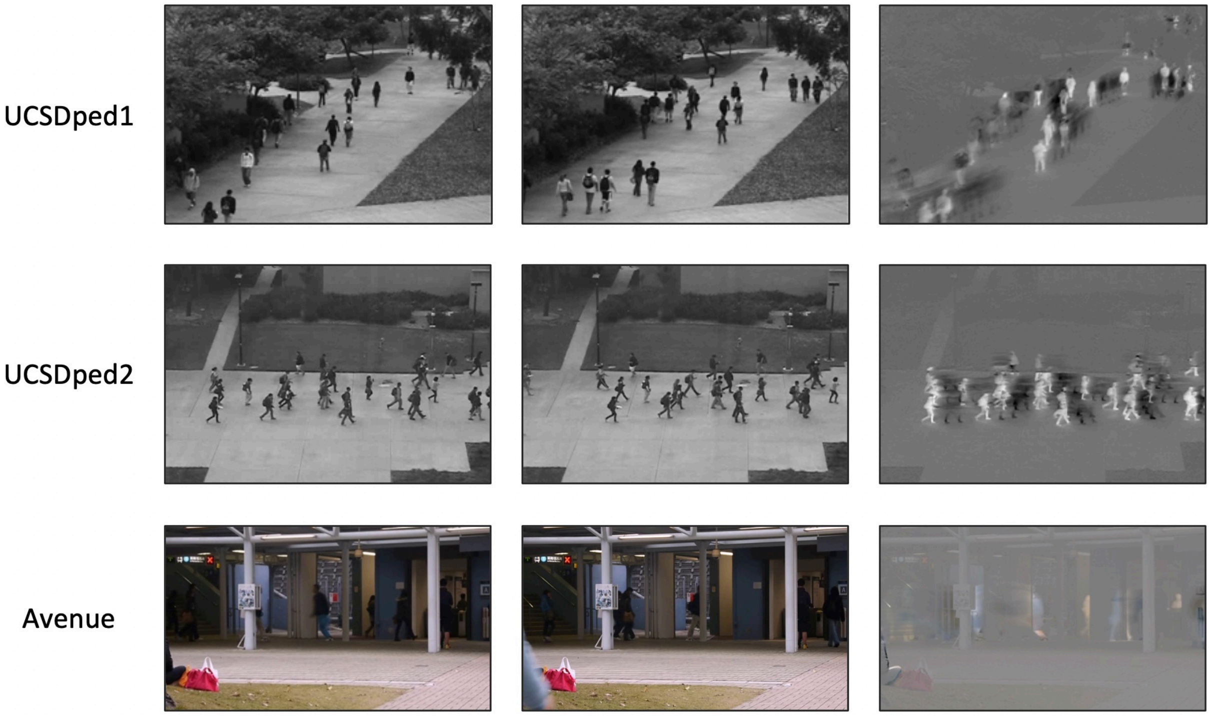


Figure 63: Three sample images from different benchmark datasets: each sample consists of frames at time t and another frame at time $t + 50$ (first and second column). The third column is the result of the dynamic image representation.

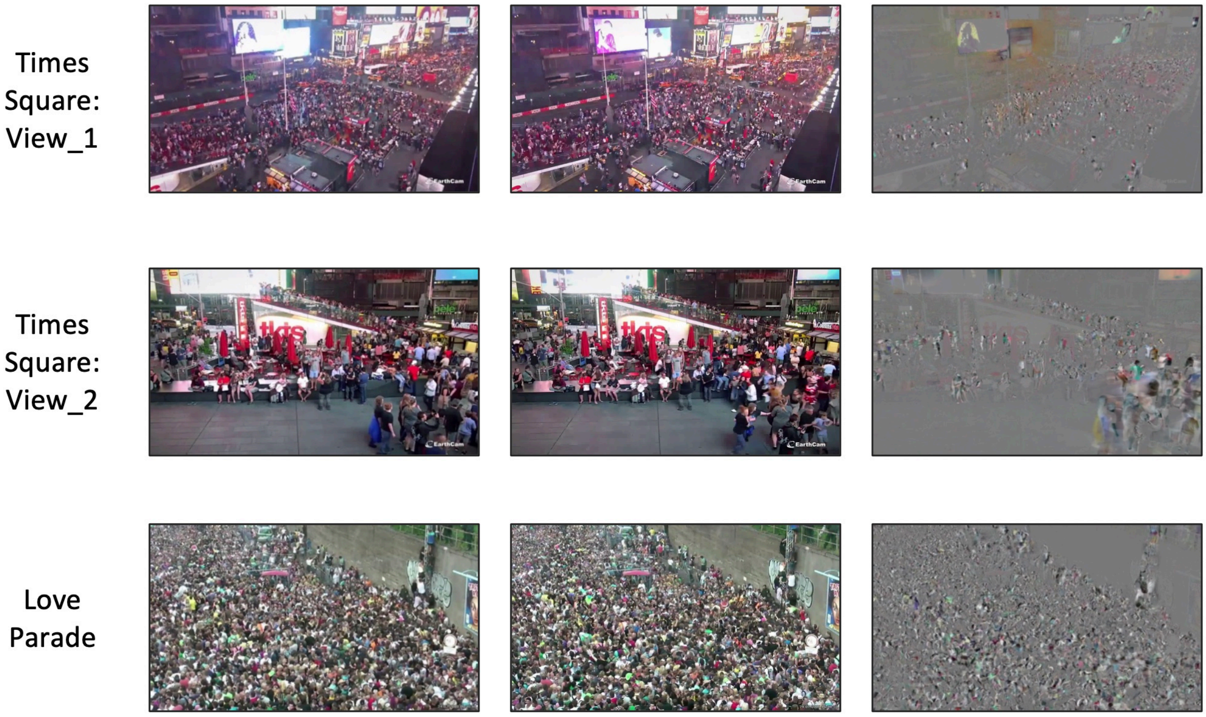


Figure 64: Three sample images from different scenes in the AHDCrowd dataset. Each sample consists of frames at time t and another frame at time $t + 50$ (first and second column) and the third column is the result of the dynamic image representation.

6.2.4 Dynamic Optical Flow

In addition to the previously suggested temporal development methods utilised as a replacement to the standard Brox optical flow, dynamic optical flow extraction was also considered. This method combines the two temporal development approaches; optical flow and dynamic images. As previously shown in Section 3.4, dynamic optical flow achieved better performance results compared to optical flow and dynamic images in the field of action recognition. Similar to the previous experiments, this method was applied to sample frames from the benchmark datasets UCSD Ped-1, UCSD Ped-2 and Avenue to illustrate the dynamic optical flow representation with $t = 10$. The method was also applied to three scenes from the AHDCrowd dataset (Mahmoud and Arafa, 2020), the results for are shown in Figures 65 and 66.

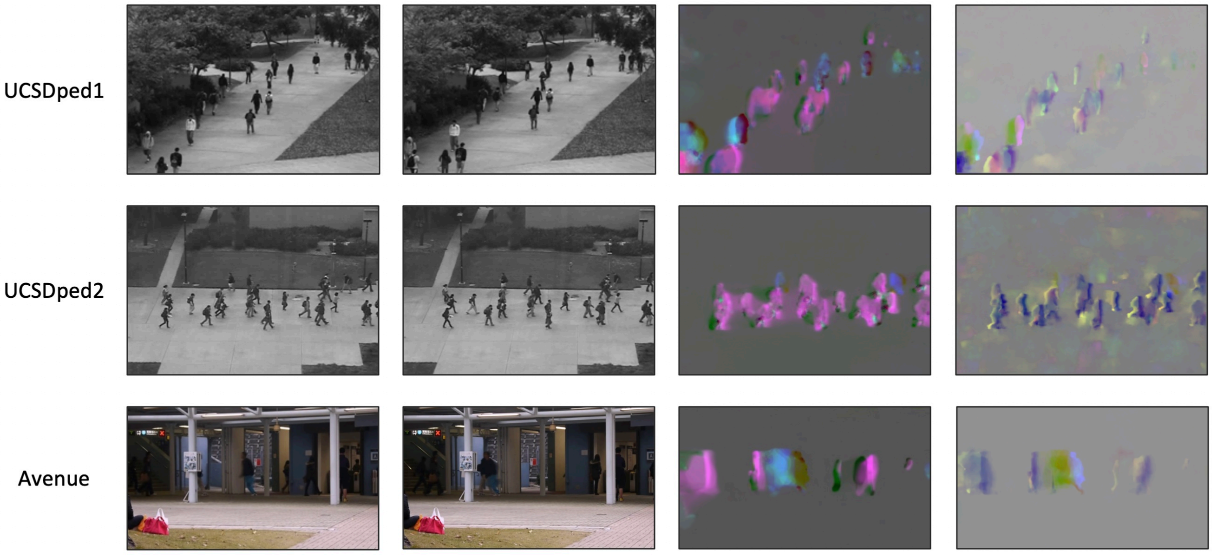


Figure 65: Three sample images from different benchmark datasets: each sample consists of frames at time t and another frame at time $t + 10$ (first and second column). The third column is the result of the dynamic optical flow (Brox) image representation and the fourth column is the result of the dynamic optical flow (FlowNet) image representation.

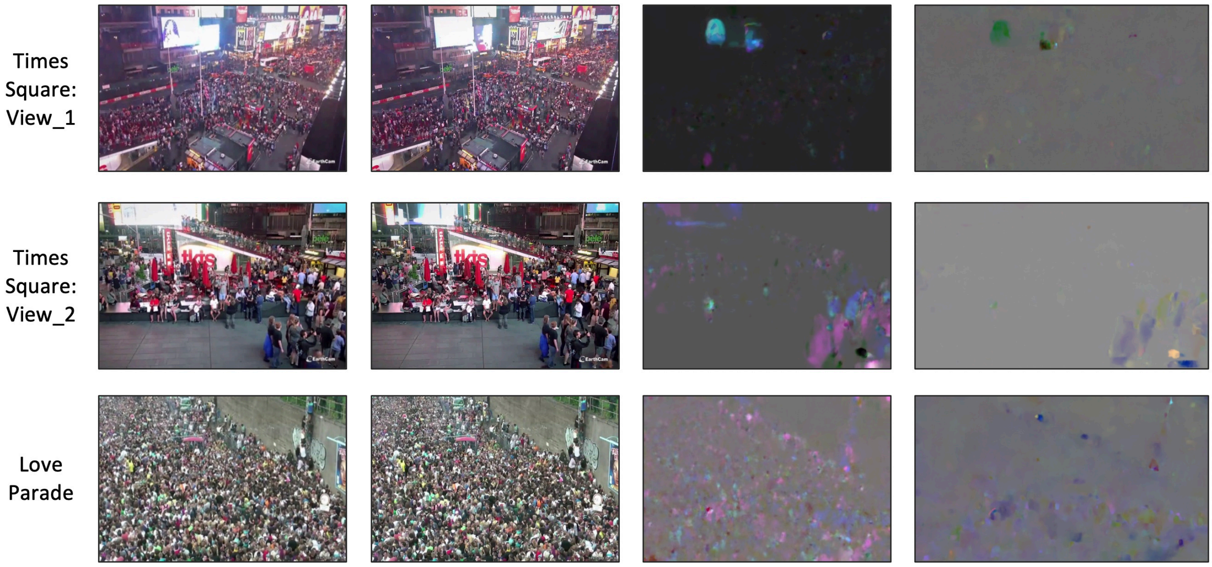


Figure 66: Three sample images from different scenes in the AHDCrowd dataset. Each sample consists of a frame at time t and another frame at time $t + 10$ (first and second column) and the third column is the result of the dynamic optical flow (Brox) image representation. The fourth column is the result of the dynamic optical flow (FlowNet) image representation.

The results produced from the application of FlowNet optical flow, dynamic images and dynamic optical flow on benchmark datasets as well as scenes from the AHDCrowd dataset will be used for the next set of crowd anomaly detection experiments using the proposed framework in

Section 4.3. Below are the combined images of the qualitative results produced from the experiments applied in the section above for easier viewing. Figures 67 and 68 show the results when applied on benchmark and the AHDCrowd datasets respectively.

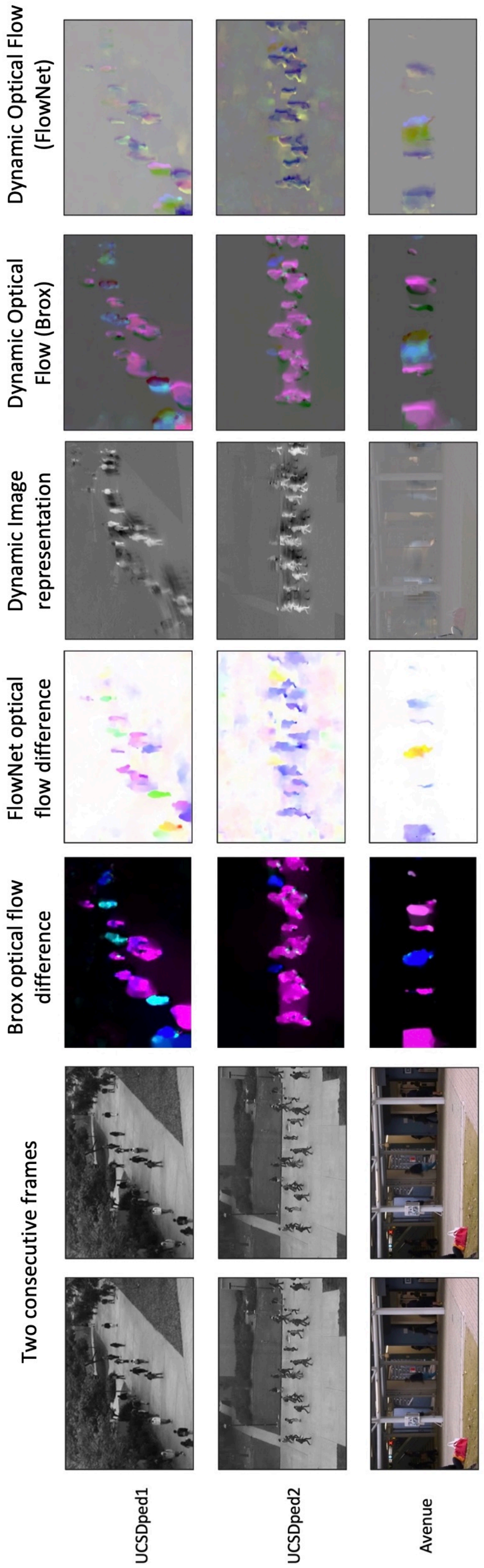

Two consecutive frames
Brox optical flow difference
FlowNet optical flow difference
Dynamic Image representation
Dynamic Optical Flow (Brox)
Dynamic Optical Flow (FlowNet)
UCSDped1
UCSDped2
Avenue

Figure 67: Qualitative results of the motion representation methods applied to benchmark datasets.

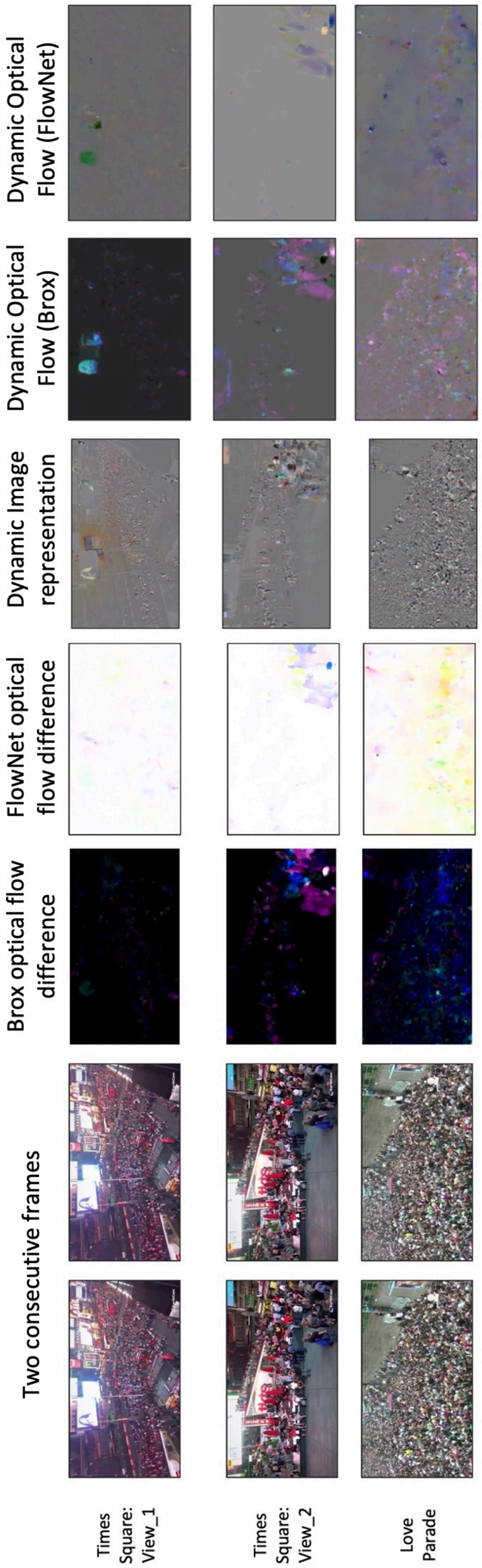

Two consecutive frames
Brox optical flow difference
FlowNet optical flow difference
Dynamic Image representation
Dynamic Optical Flow (Brox)
Dynamic Optical Flow (FlowNet)
Times Square: View_1
Times Square: View_2
Love Parade

Figure 68: Qualitative results of the motion representation methods applied to AHDCrowd dataset.

6.3 Crowd Anomaly Detection

In this section, the proposed framework for crowd anomaly detection is evaluated and compared to state-of-the-art methods using benchmark as well as the AHDCrowd datasets. The benchmark datasets used are UCSD Ped-1 (Chan et al., 2008), UCSD Ped-2 (Chan et al., 2008) and the Avenue dataset (Lu et al., 2013). These benchmark datasets are the most commonly used datasets by researchers in the field of crowd anomaly detection. Training and testing follow the experimental setup presented in Section 4.3, and the results are produced using the anomaly detection criteria: frame-level, pixel-level and dual-pixel level detection (further explained in Section 3.3) when feasible. The experimental settings are highlighted below and the obtained results of each experiment are noted.

6.3.1 Experimental settings

The applied experiments evaluate the effectiveness of utilising dynamic image representation for anomaly detection. In all the experiments noted below, two separate 3-layer DAEs, with a number of filters 32, 16 and 8 for each layer, are trained with a stride of 2 and $\beta = 0.8$ for 500 epochs. Additionally, the CGANs are trained for 10 epochs on a stochastic gradient descent with momentum 0.5 and the batch size is set to 1. All the training and testing frames are resized to 256 x 256. Each experiment utilises the input frames as well as their corresponding motion representation (optical flow or dynamic images). The motion representation used in the first experiment (Ours$_{DI}$) is the dynamic image representation of the original frames. The second experiment (Ours$_{FlowNet}$) uses Flownet (Sun et al., 2017) as the motion representation. Finally, the last two experiments use dynamic optical flow as the motion representation (Ours$_{DOF(Brox)}$ and Ours$_{DOF(FlowNet)}$). The dynamic image representation is extracted using the pre-computed optical flow difference (Brox optical flow and Flownet) of the input data. A sample visualisation of the framework tested on the UCSDped2 dataset is shown in Figure 69. The results are indicated using the evaluation metrics Area Under Curve (AUC), Equal Error Rate (EER) and the corresponding Receiver Operating Characteristic (ROC) is illustrated (further details of the evaluation metrics are noted in Section 3.5). The results of the four experiments as well as the state-of-the-art methods on the UCSD and Avenue datasets are shown in Tables 11 and 12 respectively.

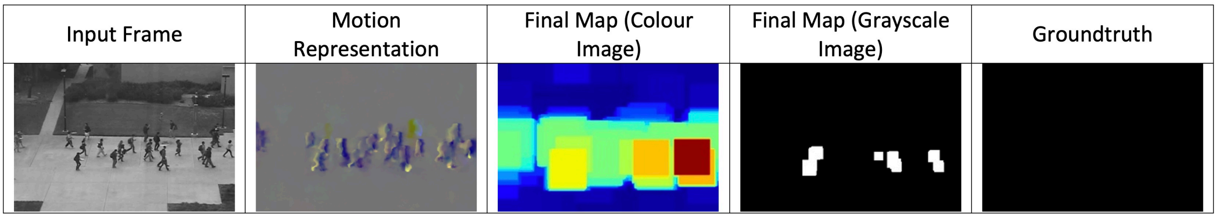


Figure 69: Sample visualisations of framework test experiment on UCSDped2.

Table 11: Comparison with the state-of-the-art on the UCSD dataset.

	Dataset							
	Ped-1				Ped-2			
Method	Frame Level		Pixel Level		Frame Level		Pixel Level	
	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)
1. Social Force	67.5	31	19.7	79	55.6	42	-	80
2. MDT	81.1	25	44.2	58	82.9	25	-	55
3. Detection at 150fps	91.8	15	63.8	43	-	-	-	-
4. Plug-and-Play	95.7	8	64.5	40.8	88.4	18	-	-
5. ConvAE	-	-	81	27.9	-	-	90	21.7
6. AMDN	92.1	16	67.2	40.1	90.8	17	90.8	17
7. GAN generative	97.4	8	70.3	35	93.5	14	-	-
8. ConvLSTM	89.9	12.5	-	-	87.4	12	-	-
9. AnoPred	83.1	-	-	-	95.4	-	-	-
10. Deep-anomaly	-	-	-	-	-	11	-	15
11. MLAD	82.34	23.5	66.6	**22.65**	97.52	4.68	94.45	4.58
12. GAN discriminative	96.8	7	70.8	34	95.5	11	-	-
13. Gaussian Mixture	94.9	11.3	71.4	36.3	92.2	12.6	78.2	19.2
Ours$_{DI}$	73.14	27.90	22.24	42.11	72.62	35.00	0	50.00
Ours$_{FlowNet}$	73.77	23.50	69.87	23.50	98.42	**3.00**	**96.38**	**3.00**
Ours$_{DOF(Brox)}$	76.96	28.95	54.60	34.74	94.72	8.00	92.42	6.0
Ours$_{DOF(FlowNet)}$	71.10	33.68	63.80	33.68	**98.84**	4.00	74.32	20.00
14. G2D	-	-	-	-	-	11	-	-
15. AEP	**97.92**	**6.07**	**74.83**	31.06	97.31	7.52	-	-

The results noted the table above are from the methods by 1. (Mehran et al., 2009), 2. (Mahadevan et al., 2010), 3. (Lu et al., 2013), 4. (Ravanbakhsh et al., 2016), 5. (Hasan et al., 2016), 6. (Xu et al., 2017), 7. (Ravanbakhsh et al., 2017), 8. (Chong and Tay, 2017), 9. (Liu et al., 2018b), 10. (Sabokrou et al., 2018), 11. (Vu et al., 2019), 12. (Ravanbakhsh et al., 2019), 13. (Fan et al., 2020), 14. (Pourreza et al., 2021a), 15. (Yu et al., 2021). The best results are indicated with bold lettering and the next best results are underlined.

Table 12: Comparison with the state-of-the-art on the Avenue dataset.

	Frame Level	
Method	AUC (↑)	EER (↓)
3. Detection at 150fps	80.5	-
5. ConvAE	70.2	25.1
8. ConvLSTM	80.3	20.7
9. AnoPred	84.9	-
11. MLAD	71.54	36.38
13. Gaussian Mixture	83.4	22.7
Ours$_{DI}$	59.00	33.00
Ours$_{FlowNet}$	85.65	13.64
Ours$_{DOF(Brox)}$	81.37	**7.00**
Ours$_{DOF(FlowNet)}$	87.38	17.00
15. AEP	**90.2**	10.07

6.3.2 Anomaly detection using Dynamic Images

As previously illustrated in Section 6.2.3, this experiment utilises the extraction of dynamic images from the input data. The extracted dynamic images are used as the motion representations given to the proposed crowd anomaly detection framework. Figure 70 shows sample images from the Avenue, UCSD Ped-1 and UCSD Ped-2 dataset produced by the DAEs. The first column illustrates a sample image corrupted with noise and its corresponding reconstructed version. The second column illustrates the dynamic image corresponding to the sample image also corrupted with noise and its reconstructed version.

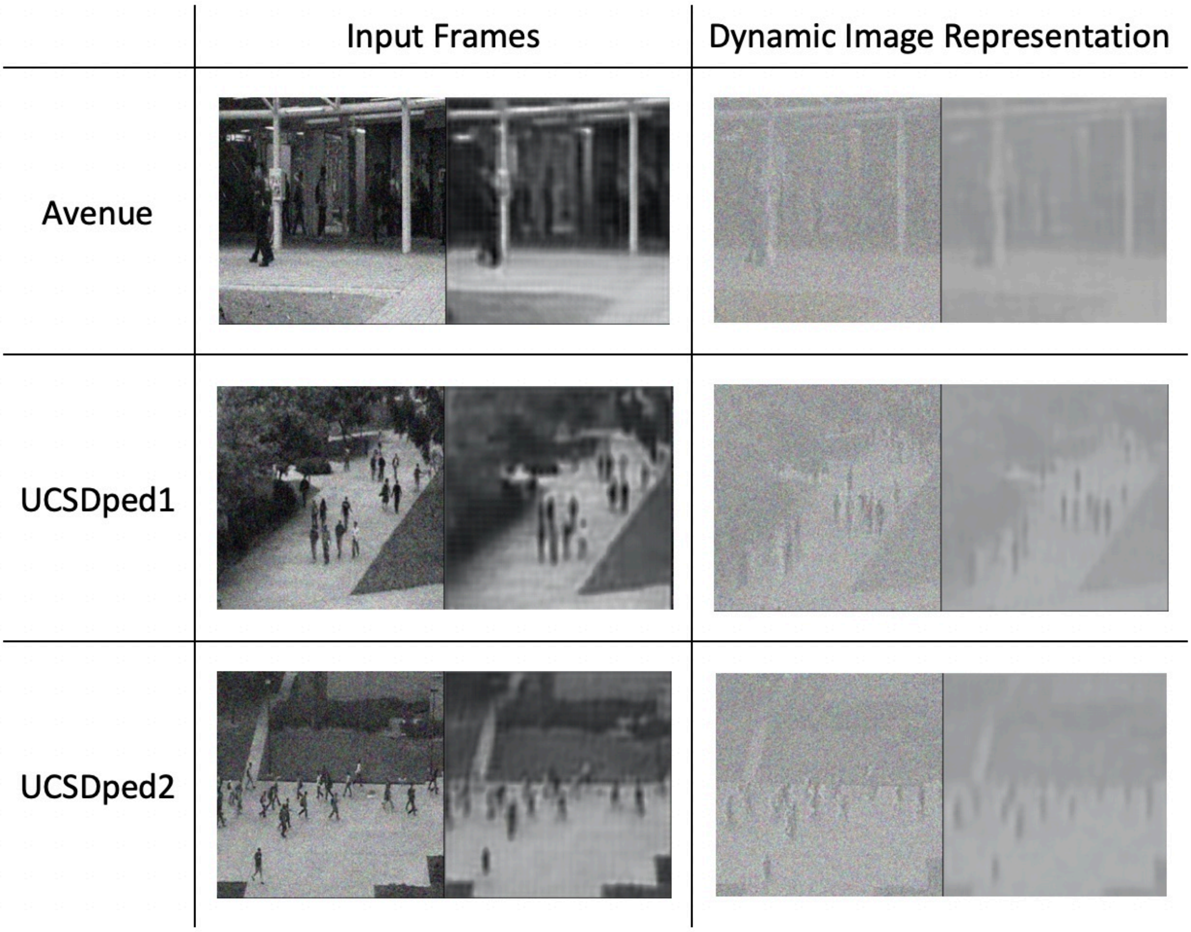


Figure 70: DAE reconstruction sample images of input frames (left) and dynamic image representation (right) from Avenue, UCSD Ped-1 and UCSD Ped-2 datasets.

The results of training and testing the proposed network with dynamic images on the Avenue, UCSD Ped-1 and UCSD Ped-2 datasets are illustrated in Figures 71, 72 and 73 respectively. The results are also noted in Tables 12 and 11 indicated as (Ours$_{DI}$). In addition to the noted results, the results of this experiment using dual-pixel detection are AUC results of 1.9% (Avenue), 0% (UCSD Ped-1) and 2.0% (UCSD Ped-2). These results indicate a significant decline in performance in comparison to the state-of-the-art. The ability to reconstruct images from their corresponding dynamic image representations does not succeed. The reconstructed data shows instances of anomalies that do not coincide with the ground-truth data. Additionally, the locations of the detected anomalies are not accurately detected which corresponds to the results achieved from pixel-level and dual-pixel level detection.

Image-to-image translation (Isola et al., 2017) using CGANs demonstrate the ability to translate edge maps or label maps to synthesised output images and the state-of-the-art crowd anomaly detection methods using CGANs (Ravanbakhsh et al., 2017; Ravanbakhsh et al., 2019; Vu et al., 2019) utilise Brox optical flow as the motion representation to be translated into an output image.

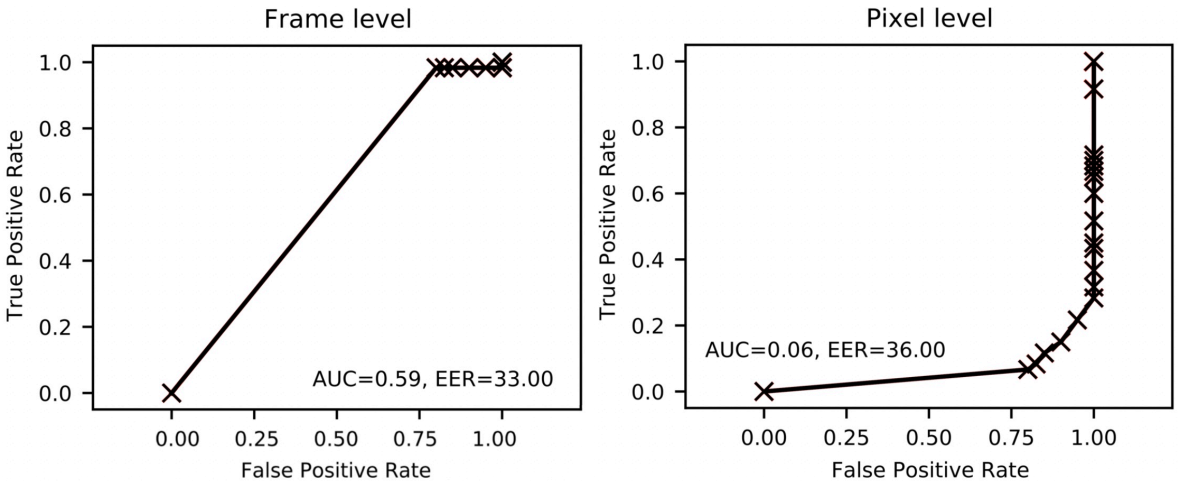


Figure 71: Ours$_{DI}$: frame-level and pixel-level ROC curves on Avenue dataset.

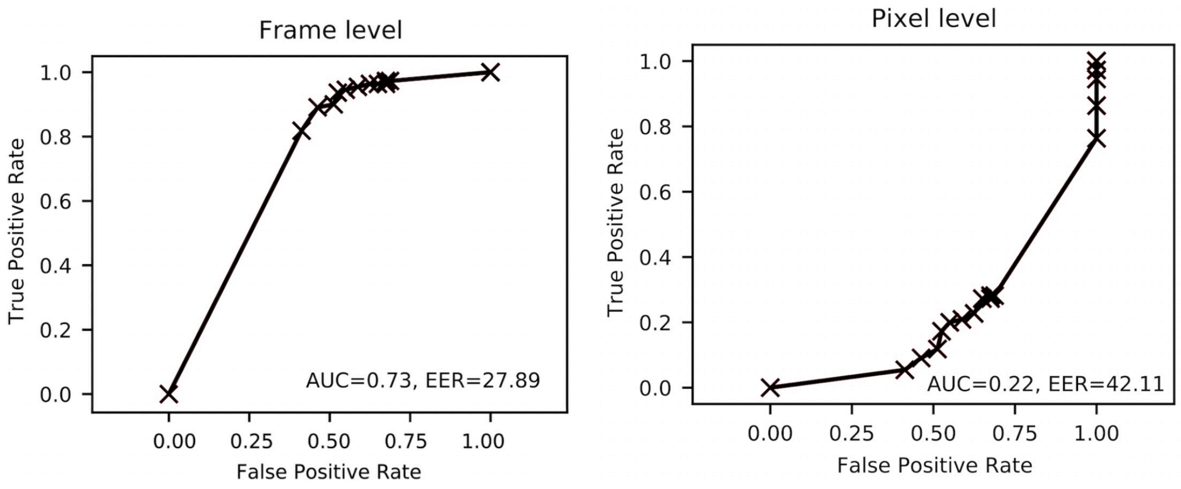


Figure 72: Ours$_{DI}$: frame-level and pixel-level ROC curves on UCSD Ped-1 dataset.

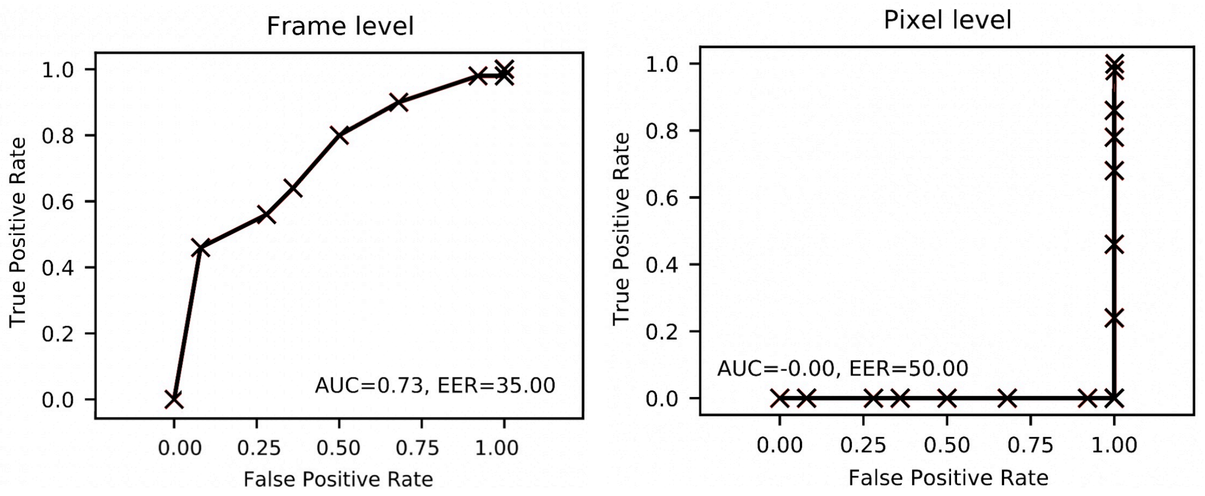


Figure 73: Ours$_{DI}$: frame-level and pixel-level ROC curves on UCSD Ped-2 dataset.

The next experiment utilises FlowNet, a novel method to compute optical flow difference, as the motion representation for the proposed framework.

6.3.3 Anomaly detection using FlowNet Optical Flow

State-of-the-art crowd anomaly detection methods using GANs have utilised Brox optical flow Brox et al. (2004) to extract motion representations. However, this experiment makes use of a more novel method for optical flow computation. As illustrated in Section 6.2, FlowNet optical flow is used to calculate the temporal development between two consecutive frames for the input data. Similar to the previous experiment, the same benchmark datasets are used to evaluate the performance of utilising FlowNet in the proposed framework. Figure 74 shows sample images from the Avenue, UCSD Ped-1 and UCSD Ped-2 dataset produced by the trained DAEs. The first column illustrates a sample image corrupted with noise and its corresponding reconstructed version. The second column illustrates the FlowNet optical flow difference corresponding to the sample image also corrupted with noise and its reconstructed version.

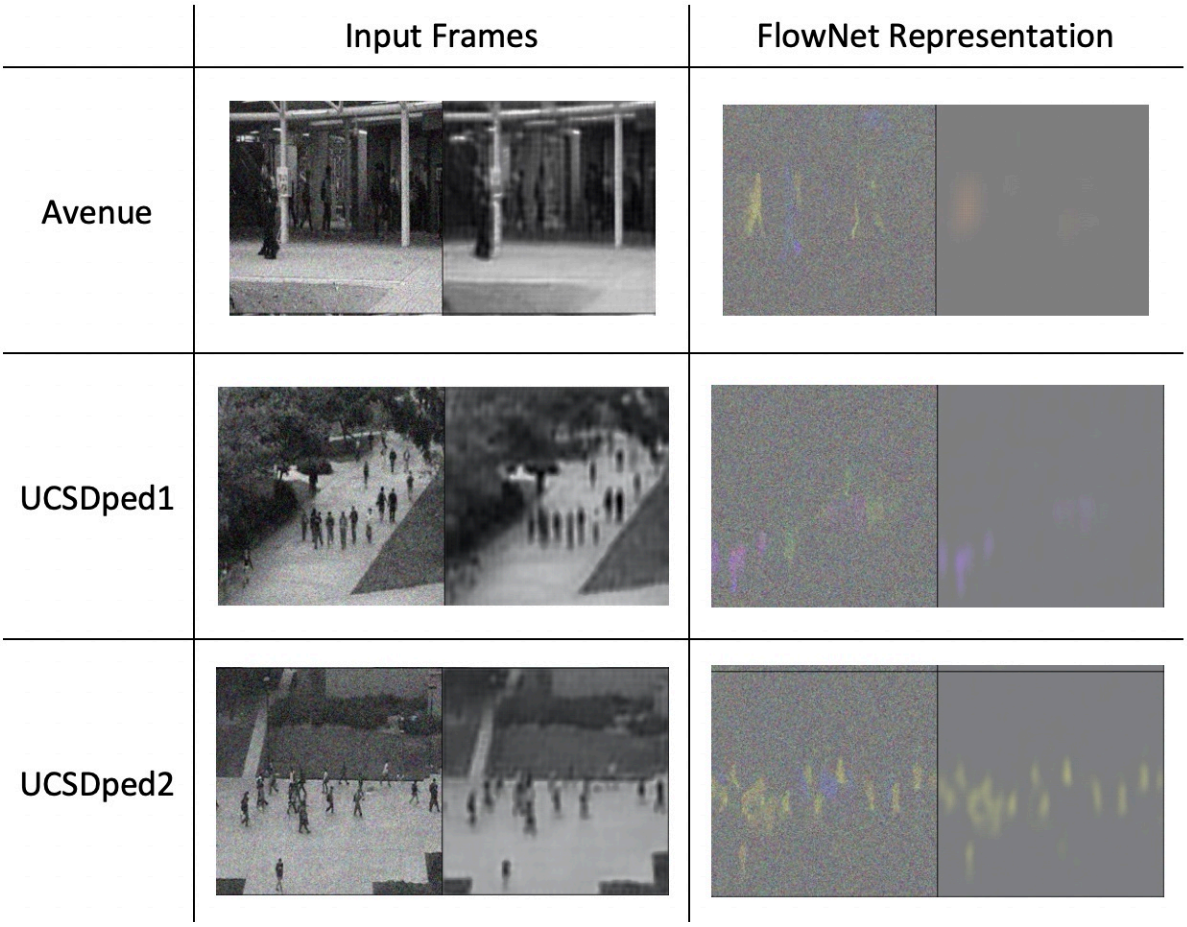


Figure 74: DAE reconstruction sample images of input frames (left) and FlowNet representation (right) from Avenue, UCSD Ped-1 and UCSD Ped-2 datasets.

AUC, EER and ROC results produced by training and testing the proposed framework on the benchmark datasets Avenue, UCSD Ped-1 and UCSD Ped-2 are shown in Figures 75, 76 and 77 respectively. The results are also shown in Tables 12 and 11 and indicated as (Ours$_{FlowNet}$). The dual-pixel detection results of this experiment are as follows 60.61% on UCSD Ped-1 and 96.37% on UCSD Ped-2. Dual-pixel detection has not been frequently applied by previous researchers but the research by Vu et al. (2019) have documented their results as follows: 60.79% on UCSD Ped-1 and 93.99% on UCSD Ped-2. In comparison to their method, this experiment has demonstrated a 2.38% improvement on the UCSD Ped-2 dataset and comparable results on UCSD Ped-1. Additionally, as shown in Tables 12 and 11 this experiment shows a 0.9% AUC and 1.68% improvement in frame-level detection on the UCSD Ped-2 dataset in comparison to the best-achieved results by state-of-the-art. In addition to these improvements, pixel-level detection AUC and EER results show a 1.93% and 1.58% improvement on the UCSD Ped-2 dataset. However, the frame-level detection results on the UCSD Ped-1 dataset show lower performance in comparison to the other methods. This is due to the ground-truth data of the UCSD Ped-1 dataset being mislabelled as discovered by Vu et al. (2019). The pixel-level detection results from this experiment on the UCSD Ped-1 dataset are comparable to other methods demonstrating the effectiveness of anomaly localisation. Frame-level detection on the Avenue dataset has also displayed effective results with at least 0.75% AUC and 7.06% EER improvement than other methods.

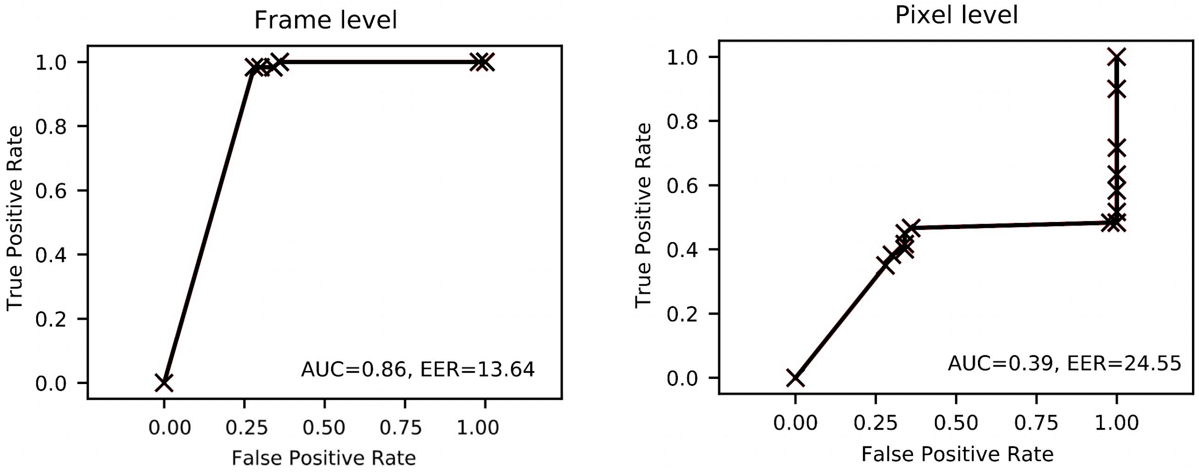


Figure 75: Ours$_{FlowNet}$: frame-level and pixel-level ROC curves on Avenue dataset.

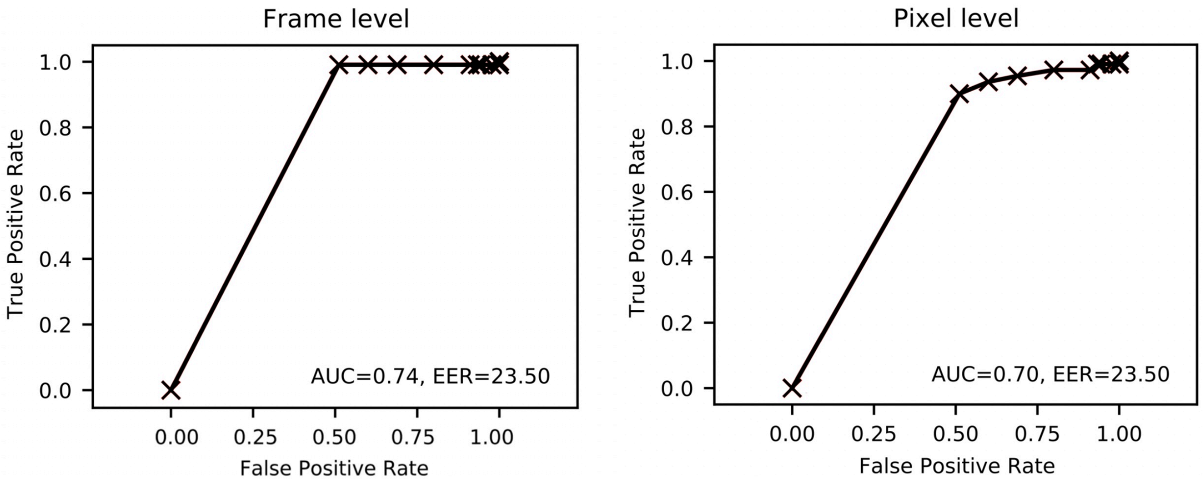


Figure 76: Ours$_{FlowNet}$: frame-level and pixel-level ROC curves on UCSD Ped-1 dataset.

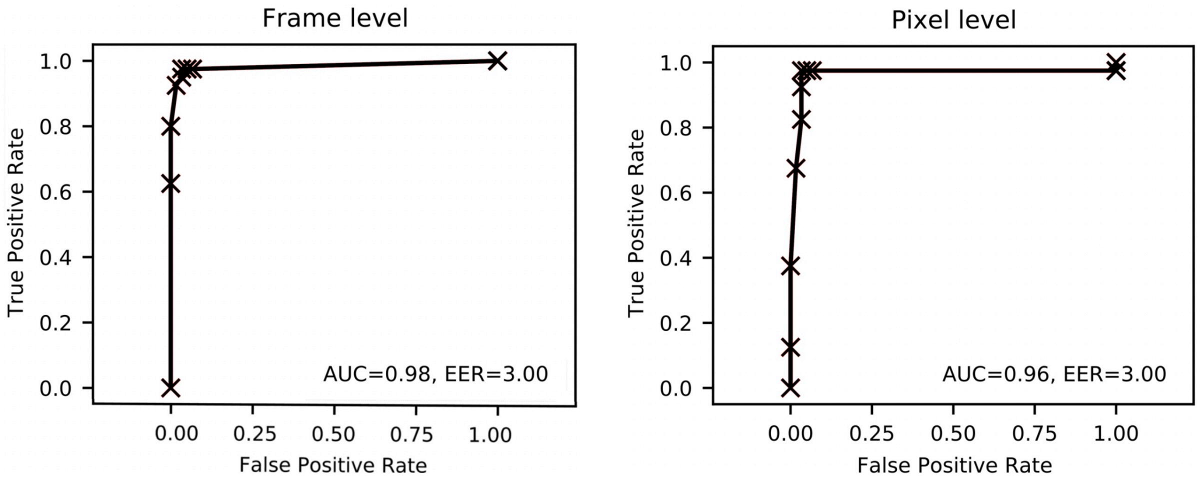


Figure 77: Ours$_{FlowNet}$: frame-level and pixel-level ROC curves on UCSD Ped-2 dataset.

The main contribution of this research is utilising Dynamic Image (Bilen et al., 2016) to extract temporal development from a set of input images represented as one motion representative image. The dynamic image motion representation is used as a substitute for optical flow difference as motion representation. However, as shown by the results from the first experiment (Section 6.3.2), the dynamic image representation of the raw input frames do not enhance the detection results of the framework. Consequently, dynamic optical flow extraction is utilised in the next experiments.

6.3.4 Anomaly detection using Dynamic Optical Flow

The following experiments use dynamic optical flow as motion representations, dynamic image extraction is applied to the optical flow representation of the raw input data instead of the raw data itself. Dynamic optical flow has shown better results than dynamic images in the field of action recognition as shown in Section 3.4. Therefore, for the following experiments dynamic optical flow is utilised as follows. The first experiment (Section 6.3.4.1) extracts dynamic images from the optical flow difference computed using Brox (Brox et al., 2004) as the motion representation for the proposed framework. Whereas the second experiment (Section 6.3.4.2) extracts dynamic images from the optical flow difference computed using FlowNet (Sun et al., 2017) as the motion representation in the anomaly detection method proposed.

6.3.4.1 Dynamic Brox Optical Flow

In this experiment, a dynamic image representation of the optical flow difference computed using Brox (Brox et al., 2004) is used as the motion representation data for the proposed crowd anomaly detection method. An illustration of the dynamic optical flow representation of sample images taken from benchmark datasets is shown in Section 6.2.4. The same benchmark datasets used in the previous experiment; UCSD Ped-1, UCSD Ped-2 and Avenue are used to evaluate the proposed framework. Sample images from these datasets and their corresponding reconstructed version are illustrated in Figure 78. The first column displays the corrupted sample image and the reconstructed version produced by the DAE. The second column shows the corrupted dynamic optical flow (Brox) and the version reconstructed by the DAE.

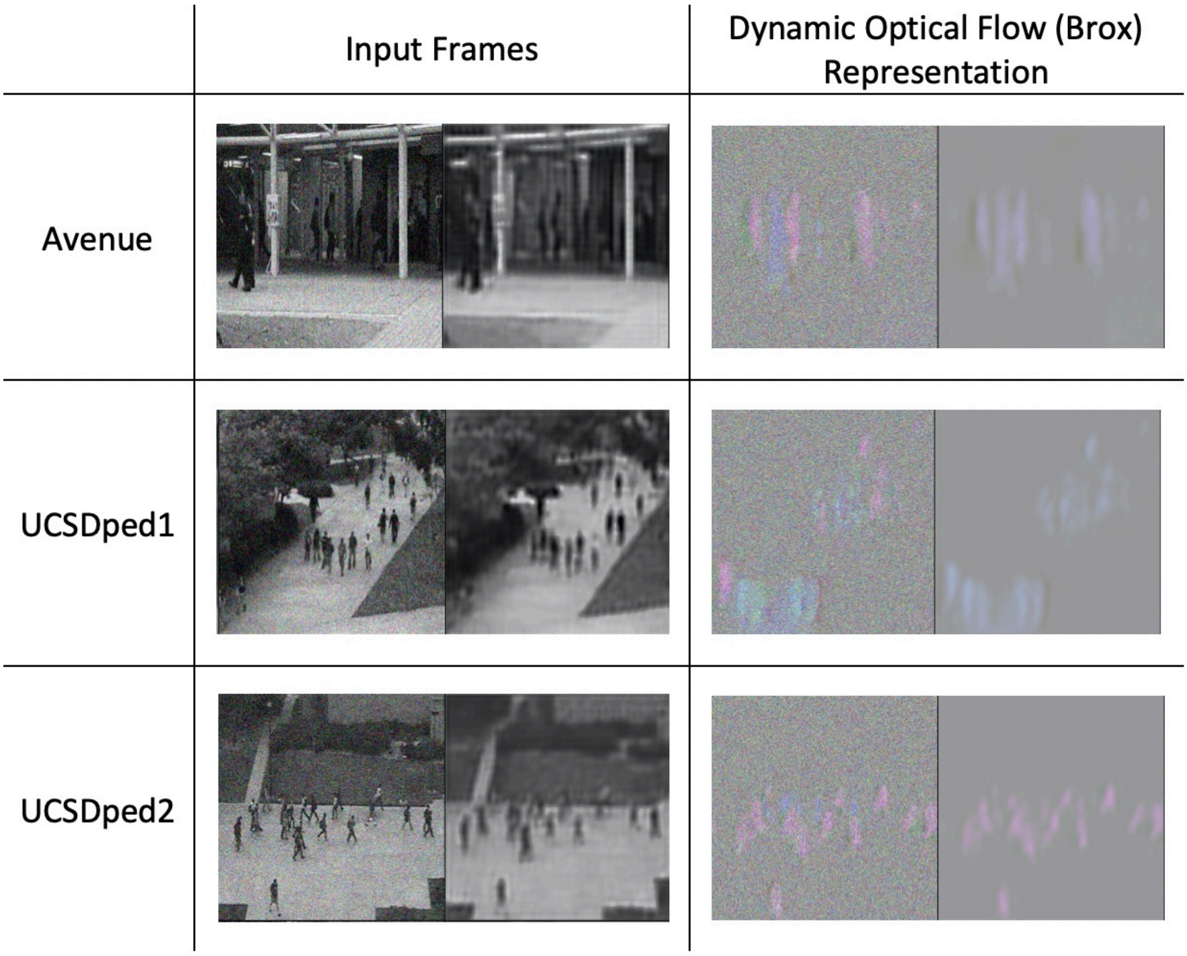


Figure 78: DAE reconstruction sample images of input frames (left) and dynamic optical flow (Brox) image representation (right) from Avenue, UCSD Ped-1 and UCSD Ped-2 datasets.

The detection results produced by utilising dynamic optical flow (Brox) as the motion representation for the proposed framework are shown in Figures 79, 80 and 81 for the Avenue, UCSD Ped-1 and UCSD Ped-2 respectively. Frame-level and pixel-level detection results, indicated as (Ours$_{DOF(Brox)}$), are also documented in Tables 12 and 11. Although the results do not show an improvement in performance in comparison to the other anomaly detection method documented applied to UCSD Ped-1 and UCSD Ped-2, the results are competitive. However, the application of this method to the Avenue dataset shows a 13.70% EER improvement and AUC results that are comparable to the other methods. It is noted that the frame-level detection results variance between this experiment and other methods is bigger than the variance in pixel-level detection. This indicates the ability of this experiment to detect anomalies on pixel-level surpasses its ability to detect anomalies on frame-level.

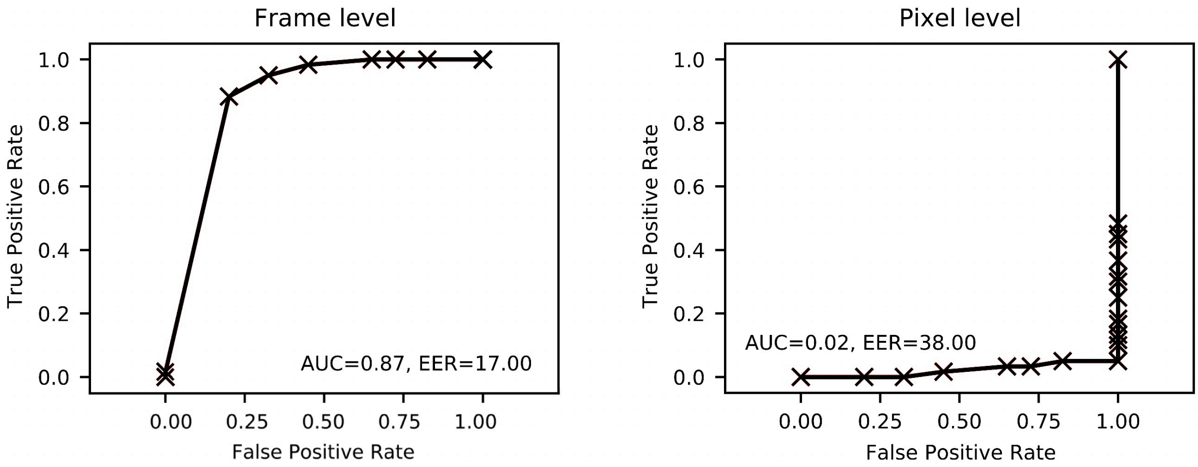


Figure 79: Ours$_{DOF(Brox)}$: frame-level and pixel-level ROC curves on Avenue dataset.

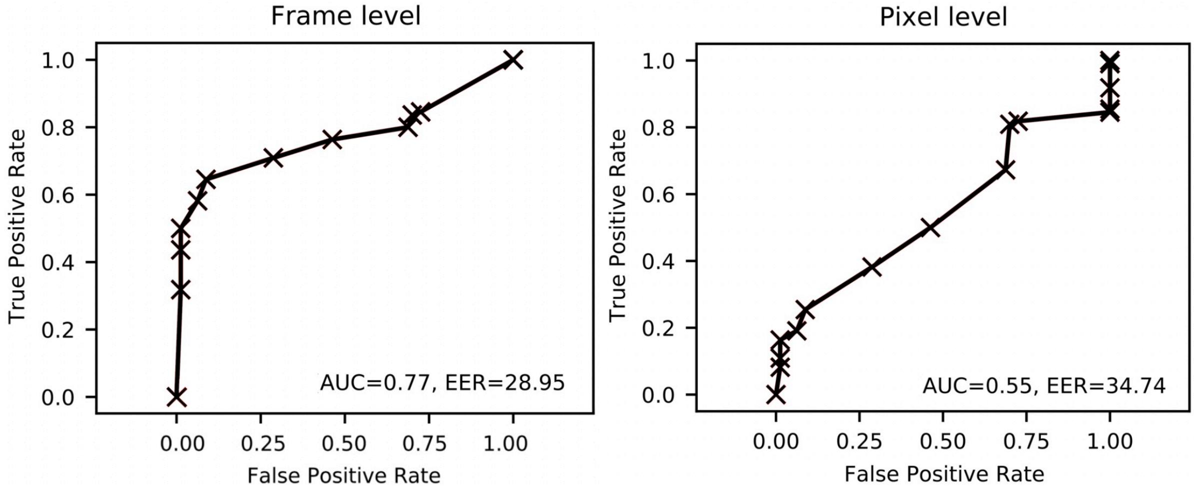


Figure 80: Ours$_{DOF(Brox)}$: frame-level and pixel-level ROC curves on UCSD Ped-1 dataset.

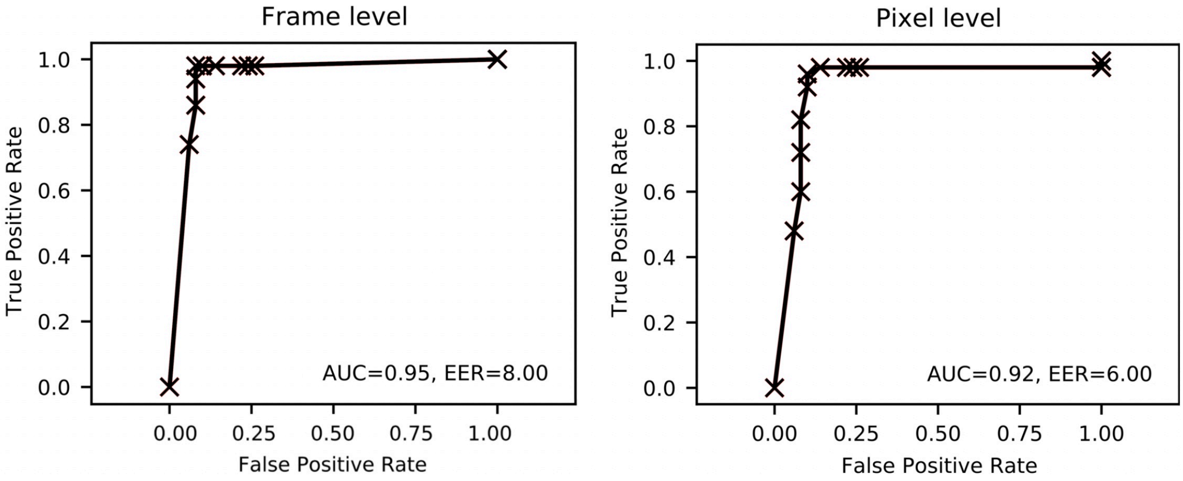


Figure 81: Ours$_{DOF(Brox)}$: frame-level and pixel-level ROC curves on UCSD Ped-2 dataset.

The last experiment is applied by extracting the dynamic image representations of the optical flow difference computed using FlowNet. As demonstrated in the second experiment, the use of FlowNet for optical flow computation has enhanced the ability of the proposed framework in the detection of anomalies on frame-level, pixel-level and dual-pixel level. Therefore, the next experiment uses dynamic optical flow (FlowNet) as the motion representation given to the proposed anomaly detection method.

6.3.4.2 Dynamic FlowNet Optical Flow

In this experiment, a dynamic image representation of the optical flow difference computed using FlowNet (Sun et al., 2017) is used as the motion representation data given to the proposed crowd anomaly detection framework. The dynamic optical flow representations of sample images taken from benchmark datasets are illustrated in Section 6.2.4. Similar to the previous experiments, training and testing were applied on the UCSD Ped-1, UCSD Ped-2 and Avenue datasets. Reconstruction samples produced from the trained DAEs are shown in Figure 82, the first column displays the sample image (corrupted with noise) and the corresponding reconstructed version. The second column shows the dynamic optical flow (FlowNet) corrupted with noise and the reconstructed version.

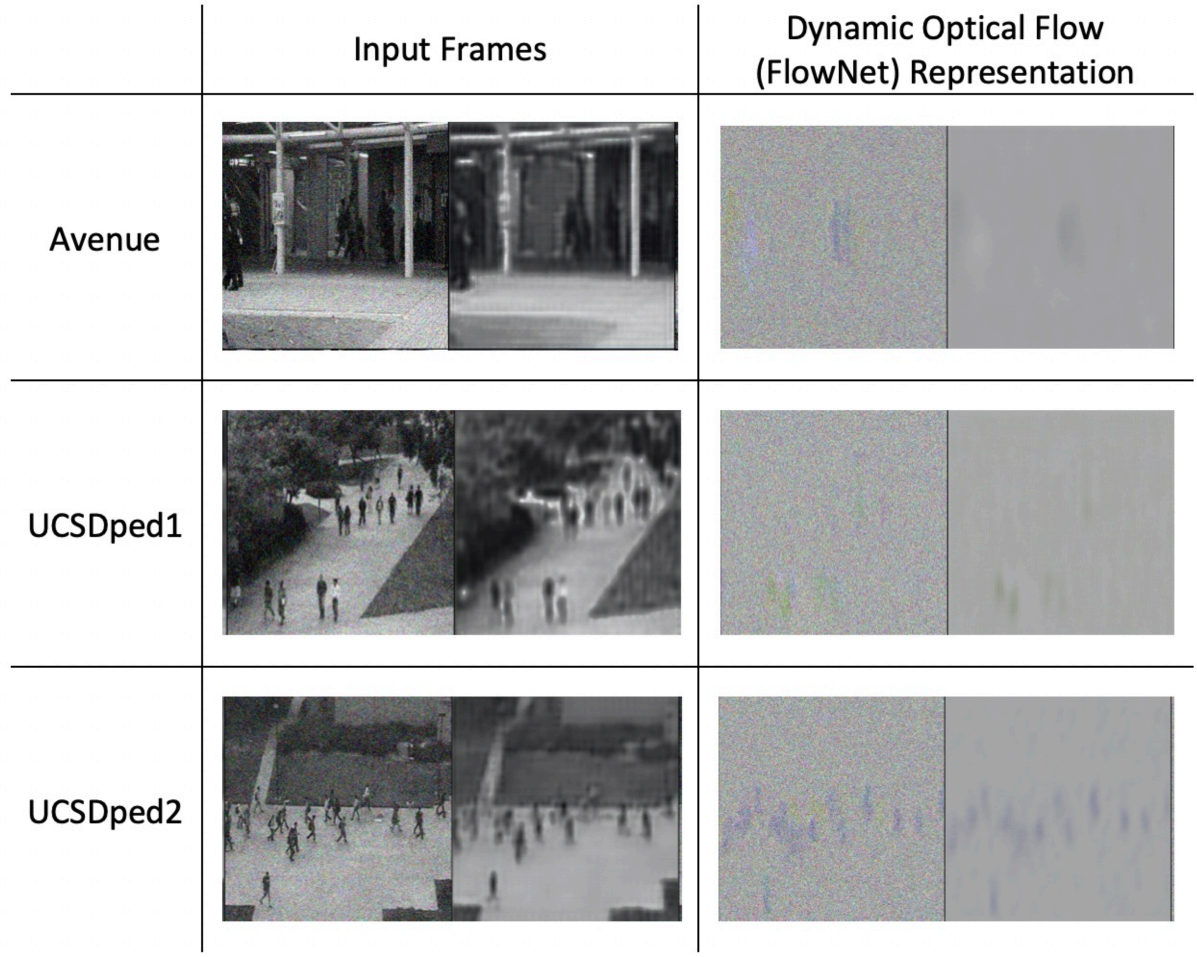


Figure 82: DAE reconstruction sample images of input frames (left) and dynamic optical flow (FlowNet) image representation (right) from Avenue, UCSD Ped-1 and UCSD Ped-2 datasets.

Detection results given from combining dynamic optical flow (FlowNet) with the proposed anomaly detection framework are shown in Figures 83, 84 and 85 for the Avenue, UCSD Ped-1 and UCSD Ped-2 respectively. The results, indicated as (Ours$_{DOF(FlowNet)}$), are also documented in Tables 12 and 11. In comparison to the state-of-the-art, the frame-level results on UCSD Ped-1 are lower, however, as previously noted this is likely caused by the mislabelling of the ground-truth data. Nevertheless, the pixel-level detection results on UCSD Ped-1 are comparable to the other methods. Additionally, the AUC and EER frame-level detection results on UCSD Ped-2 shown an improvement of 1.32% and 0.68% in comparison to other methods. The pixel-level results on UCSD Ped-2 show a decline in performance, this indicates the ineffectiveness of this experiment to accurately localise anomalies. On the other hand, the frame-level results on the Avenue dataset show a 2.48% and 3.7% improvement in AUC and EER respectively.

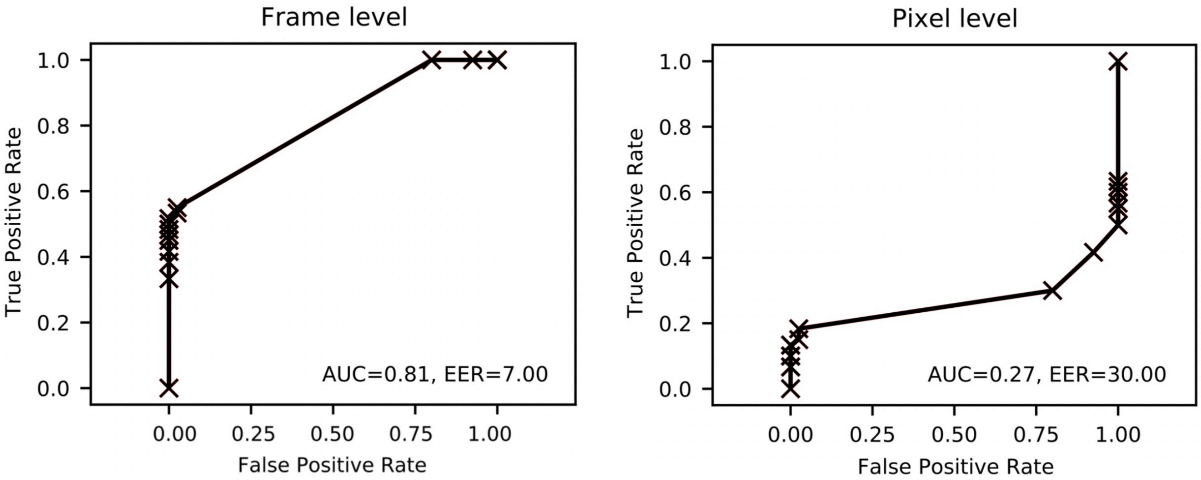


Figure 83: Ours$_{DOF(FlowNet)}$: frame-level and pixel-level ROC curves on Avenue dataset.

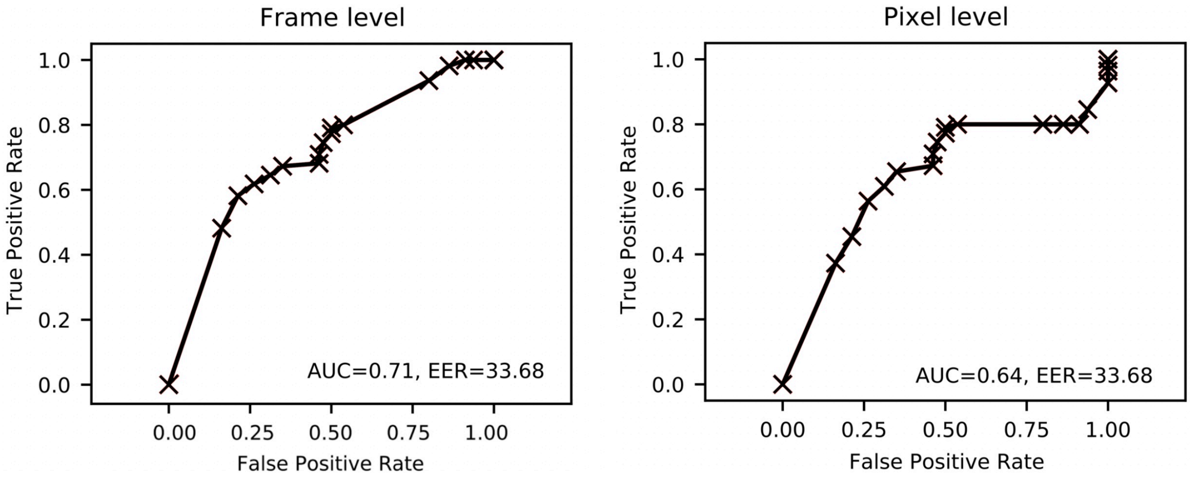


Figure 84: Ours$_{DOF(FlowNet)}$: frame-level and pixel-level ROC curves on UCSD Ped-1 dataset.

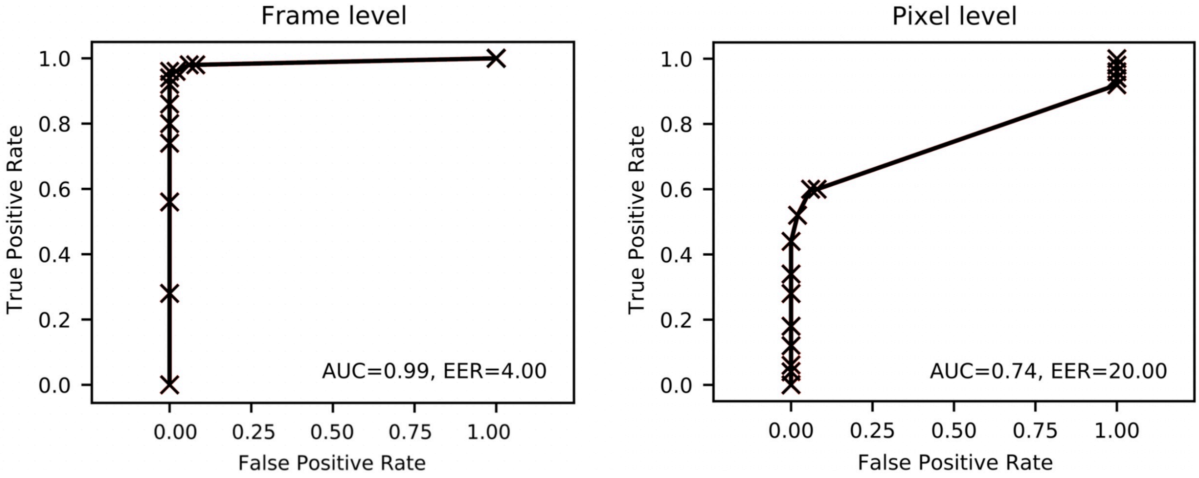


Figure 85: Ours$_{DOF(FlowNet)}$: frame-level and pixel-level ROC curves on UCSD Ped-2 dataset.

Below are ROC curves of each motion representation (Ours$_{DI}$, Ours$_{FlowNet}$, Ours$_{DOF(Brox)}$, Ours$_{DOF(FlowNet)}$) combined in a single image for easier viewing. Figures 86, 87, 88, 89, 90, 91 are the frame-level and pixel-level ROC curves applied on the Avenue, UCSD Ped-1, and UCSD Ped-2 datasets respectively.

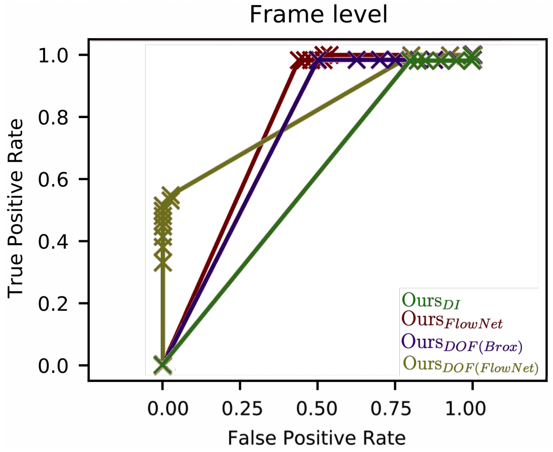


Figure 86: Frame-level ROCs on Avenue dataset.

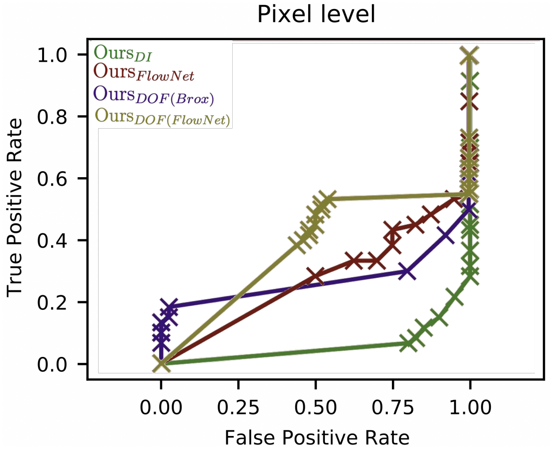


Figure 87: Pixel-level ROCs on Avenue dataset.

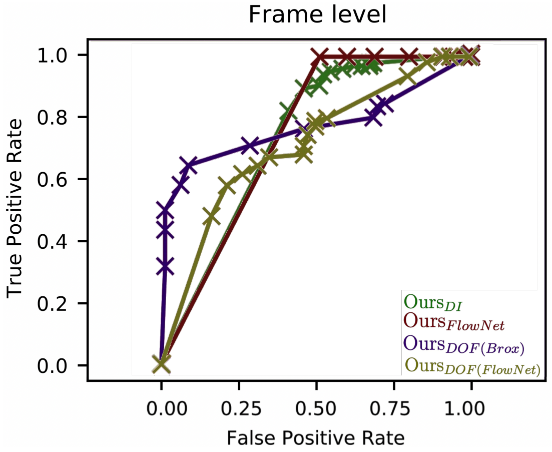


Figure 88: Frame-level ROCs on UCSD Ped-1 dataset.

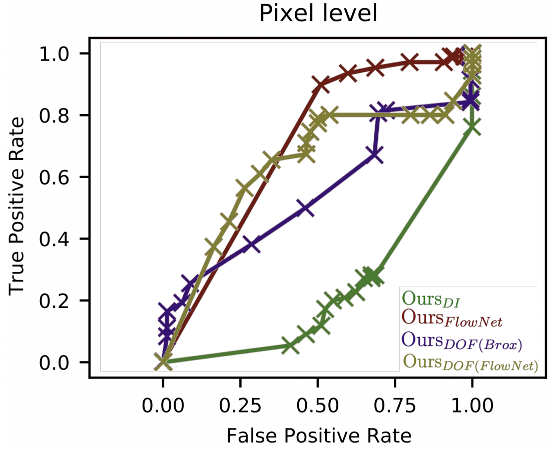


Figure 89: Pixel-level ROCs on UCSD Ped-1 dataset.

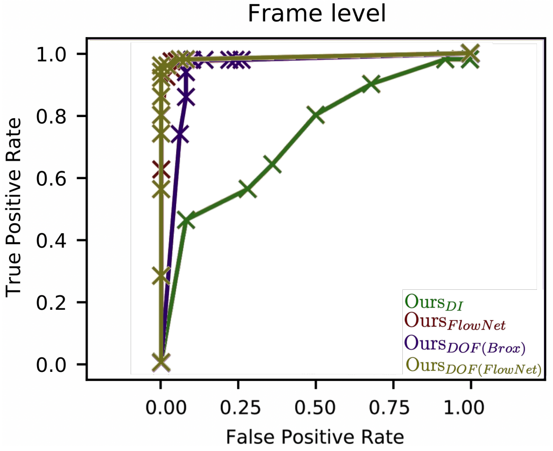


Figure 90: Frame-level ROCs on UCSD Ped-2 dataset.

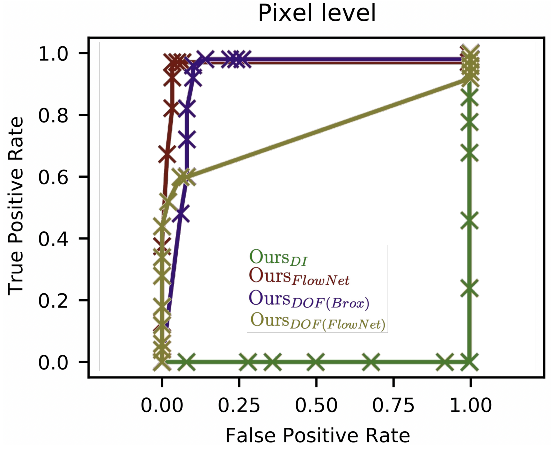


Figure 91: Pixel-level ROCs on UCSD Ped-2 dataset.

6.4 High-Density Crowd Anomaly Detection

In this section, the proposed framework for crowd anomaly detection is evaluated against the AHDCrowd dataset (described in Chapter 5). Three scenes from the dataset are used for experimentation; Times Square: View_1, Times Square: View_2 and Love Parade. Training and testing follow the experimental setup documented in Section 4.3, and the results are produced using the anomaly detection criteria; frame-level (further explained in Section 3.3). The experimental settings applied are similar to the experimental setting described in Section 6.3.1. The three experiments described below utilise input frames as well as their corresponding motion representation (optical flow or dynamic images). Four different motion representation are utilised for each experiment; dynamic image representation, Flownet optical flow (Sun et al., 2017), dynamic optical flow representations using Brox optical flow (Brox et al., 2004) and dynamic optical flow representations (using Flownet optical flow). The four motion representations are indicated as Ours$_{DI}$, Ours$_{FlowNet}$, Ours$_{DOF(Brox)}$ and Ours$_{DOF(FlowNet)}$ respectively. The results are produced using the evaluation metrics Area Under Curve (AUC), Equal Error Rate (EER) and the corresponding Receiver Operating Characteristic (ROC) is illustrated (further details of the evaluation metrics are noted in Section 3.5). The results of the experiments are shown in Tables 13, the best achieved results are indicated using bold lettering.

Table 13: Frame-level detection result using the proposed framework on three scenes from the AHDCrowd dataset.

	Times Square: View 1		Times Square: View 2		Love Parade	
Method	Frame Level		Frame Level		Frame Level	
	AUC(↑)	EER(↓)	AUC(↑)	EER(↓)	AUC(↑)	EER(↓)
Ours$_{DI}$	48.96	47.85	47.36	**19.00**	36.81	33.00
Ours$_{FlowNet}$	**65.23**	40.00	26.47	65.00	85.16	**9.00**
Ours$_{DOF(Brox)}$	64.35	**20.71**	**70.37**	31.00	**96.15**	**9.00**
Ours$_{DOF(FlowNet)}$	44.70	46.66	32.20	61.81	73.91	14.54

6.4.1 Times Square: View 1

In this experiment, the proposed framework is applied to the Times Square: View 1 scene from the AHDCrowd dataset. The scene includes footage of a high-density crowd that thought they heard gunshots and started to panic and quickly disperse. The footage is shot from an angled view. The proposed framework is trained and tested on this scene using the four motion representation previously mentioned. Figure 92 illustrates the input given to the DAEs and its corresponding output. The first column shows a sample image corrupted with noise and its corresponding reconstructed version. The second column illustrates the dynamic image representation, FlowNet representation, dynamic optical flow (Brox) representation and dynamic optical flow (FlowNet) representation corresponding to the sample image corrupted with noise. Additionally, next to each motion representation is the corresponding reconstructed version.

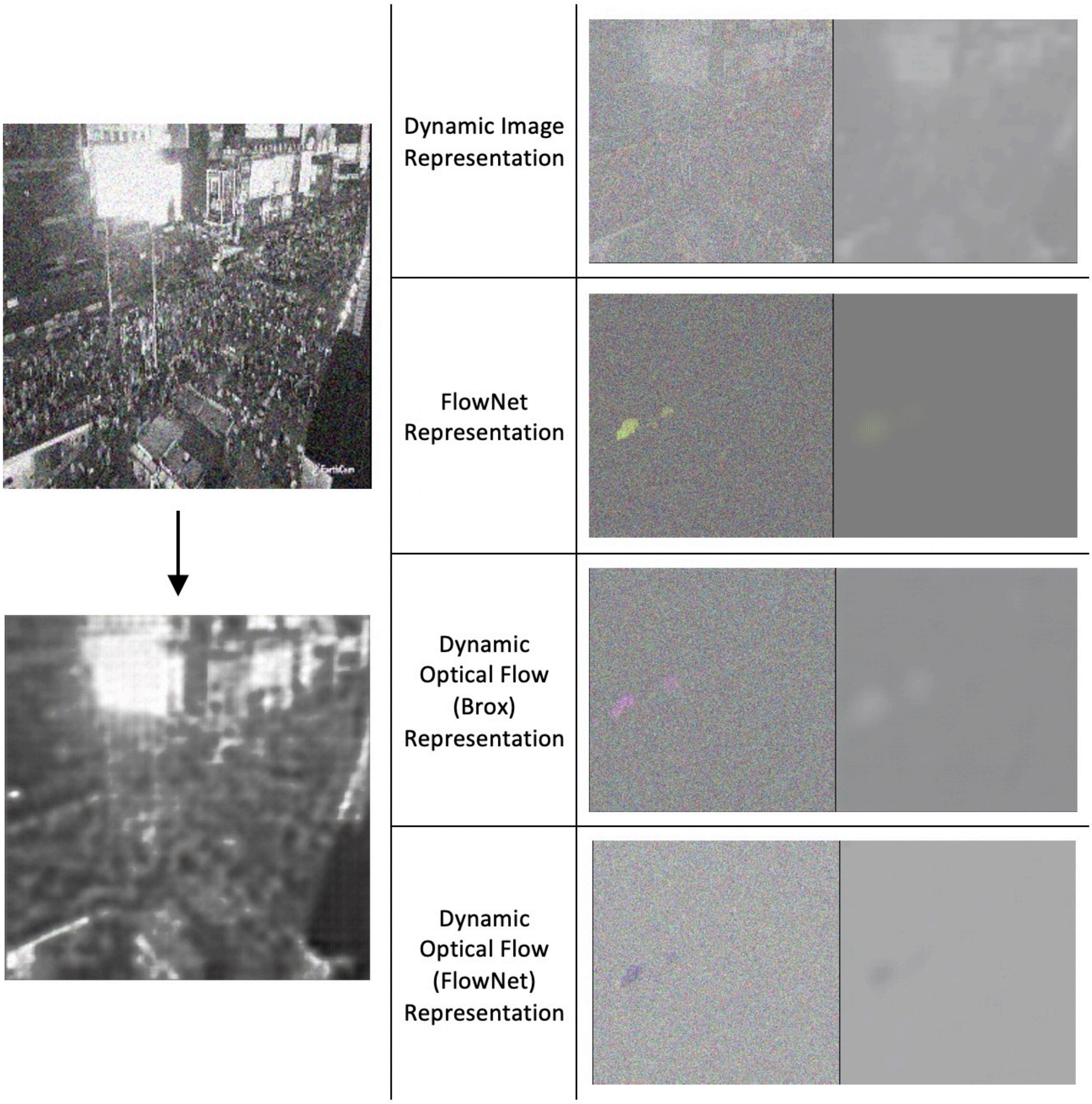

Dynamic Image Representation
FlowNet Representation
Dynamic Optical Flow (Brox) Representation
Dynamic Optical Flow (FlowNet) Representation

Figure 92: DAE reconstruction images of sample input frames (left) and the four corresponding motion representations (right) from the Times Square View 1 scene.

The frame-level results AUC, EER and ROC curve results of training and testing the proposed network on the Times Square View 1 dataset are illustrated in Figures 93, 94, 95 and 96. These Figures are results produced from utilising the motion representations: dynamic image, FlowNet, dynamic optical flow (Brox) and dynamic optical flow (FlowNet) respectively as an input to the proposed crowd anomaly detection framework. As shown in Table 13, the best achieved AUC result, 65.23, is produced by utilising FlowNet optical flow as the motion representation. Similar results are produced using dynamic optical flow (Brox), however, the AUC results produced from using dynamic image and dynamic optical flow (FlowNet) demonstrate a significant decline in AUC performance. With respect to the EER values, the best result, 20.71, is produced using dynamic optical flow (Brox) representation. The remaining EER values also demonstrate a significant decline in performance. These results indicate that the most appropriate motion representation to be used with the proposed framework is the dynamic optical flow (Brox).

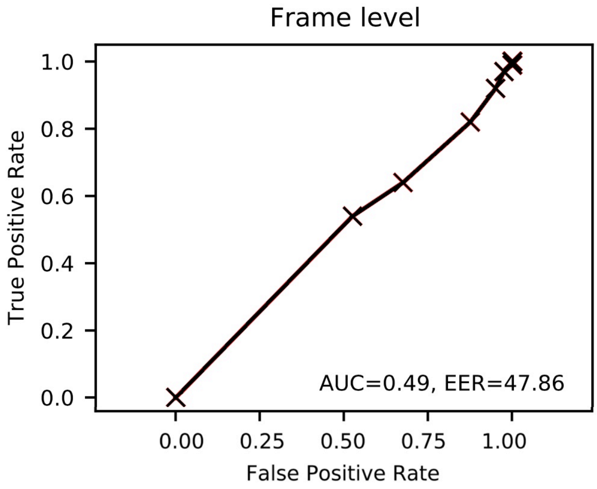


Figure 93: Ours$_{DI}$: frame-level ROC curve on Times Square View 1.

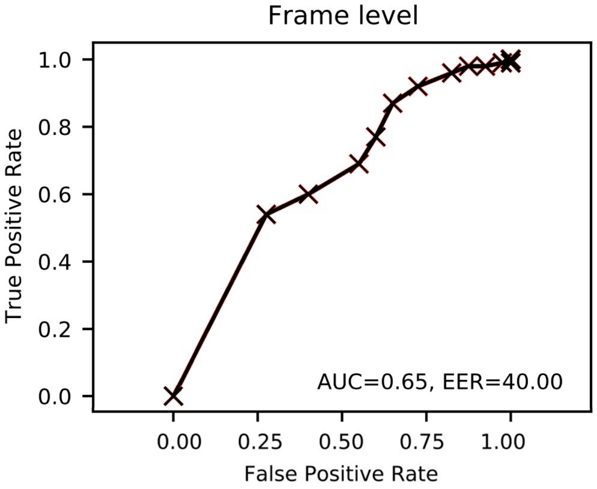


Figure 94: Ours$_{FlowNet}$: frame-level ROC curve on Times Square View 1.

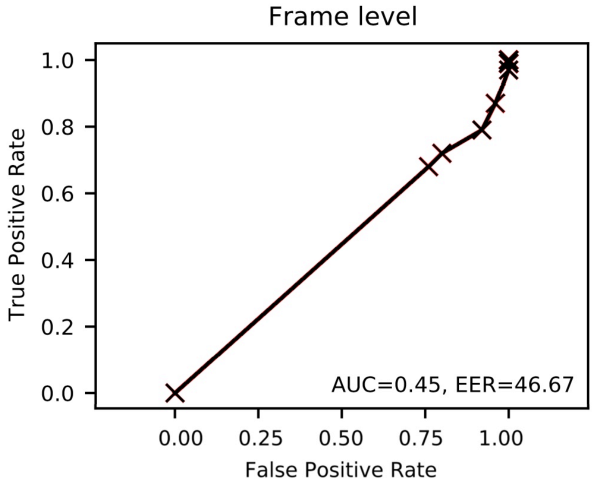


Figure 95: Ours$_{DOF(Brox)}$: frame-level ROC curve on Times Square View 1.

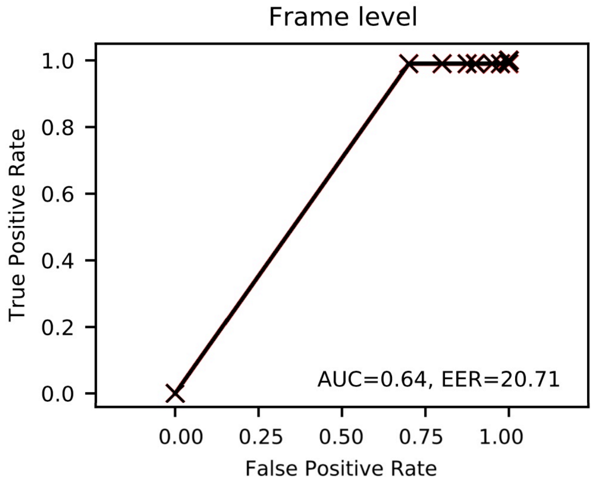


Figure 96: Ours$_{DOF(FlowNet)}$: frame-level ROC curve on Times Square View 1.

In comparison to the detection results produced by applying the Future Frame Prediction (Liu et al., 2018b) (Section 6.1.1.2.1) and Anomaly Detection Using Multilevel Representations (Vu et al., 2019) ((Section 6.1.1.3.1)) methods on this scene, the results produced from the proposed framework prove to be competitive. Details of training and testing these methods are detailed in Section 6.1. The AUC and EER values produced by training and testing Liu et al. (2018b) on this scene are 85.64 and 19.56 respectively. Additionally, the AUC and EER values produced by training and testing Vu et al. (2019) on this scene are 57.42 and 28.66 respectively. While the results produced from Liu et al. (2018b) show better performance than the proposed method, our results show a significant improvement in comparison to Vu et al. (2019).

6.4.2 Times Square: View 2

In this experiment, the proposed framework is applied to the Times Square: View 2 scene from the AHDCrowd dataset. This scene includes footage of a high-density crowd that thought they heard gunshots and started to panic and quickly disperse, unlike View 1, this footage is shot from a closeup almost eye-level shot. The crowd anomaly detection framework proposed is trained and tested on this scene using the same four motion representation as the previous experiment. Figure 97 illustrates the input given to the DAEs and its corresponding output. The first column shows a sample image corrupted with noise and its corresponding reconstructed version. The second column illustrates the dynamic image representation, FlowNet representation, dynamic optical flow (Brox) representation and dynamic optical flow (FlowNet) representation corresponding to the sample image corrupted with noise. The corresponding reconstructed version is illustrated beside each motion representation.

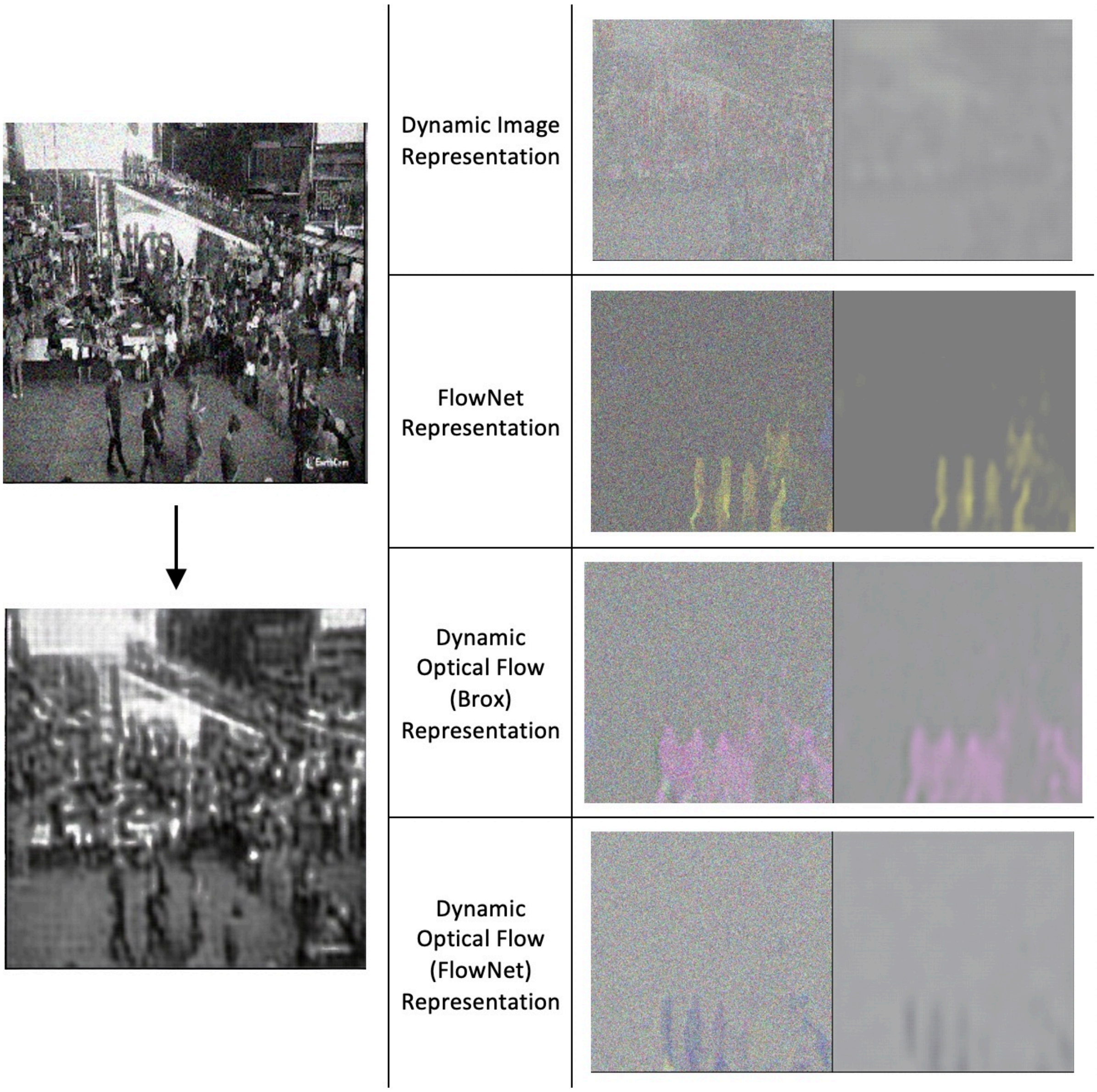

Dynamic Image Representation
FlowNet Representation
Dynamic Optical Flow (Brox) Representation
Dynamic Optical Flow (FlowNet) Representation

Figure 97: DAE reconstruction images of sample input frames (left) and the four corresponding motion representations (right) from the Times Square View 2 scene.

The frame-level results AUC, EER and ROC curve results of training and testing the proposed network on the Times Square View 2 dataset are illustrated in Figures 98, 99, 100 and 101. The figures illustrate the produced results by utilising the motion representations: dynamic image, FlowNet, dynamic optical flow (Brox) and dynamic optical flow (FlowNet) respectively as the temporal input given to the proposed crowd anomaly detection framework. As shown in Table 13, the best achieved AUC result, 70.37, is produced by using dynamic optical flow (Brox) as the motion representation. The remaining AUC results indicate that the performance of the other motion representatives are not of the same quality on this scene. With respect to the EER values, the best result, 19.00, is produced using dynamic image representations, close results are produced using dynamic optical flow (Brox) with an EER value of 31.00. The remaining EER values given from using FlowNet and dynamic optical flow (FlowNet) as motion representations indicate a significant decline in performance. Altogether, these results indicate that the most appropriate motion representation to be used with the proposed framework for this scene is the dynamic optical flow (Brox) representation.

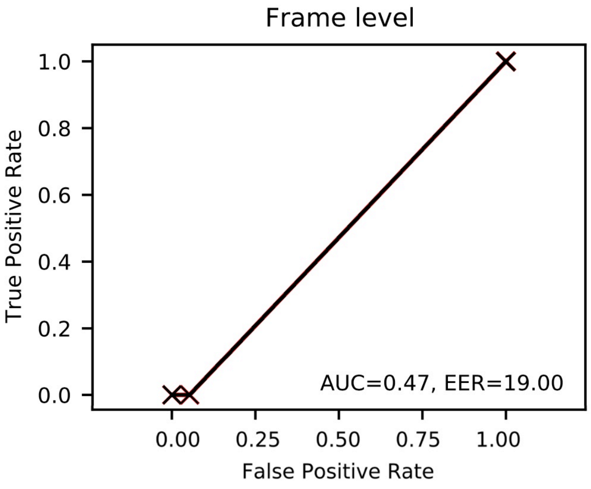


Figure 98: Ours$_{DI}$: frame-level ROC curve on Times Square View 2.

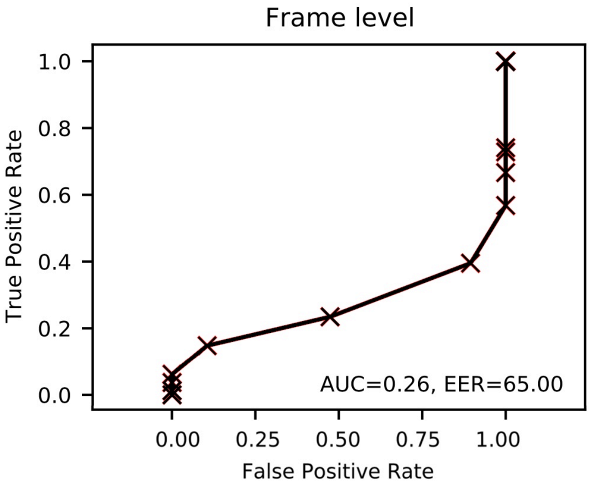


Figure 99: Ours$_{FlowNet}$: frame-level ROC curve on Times Square View 2.

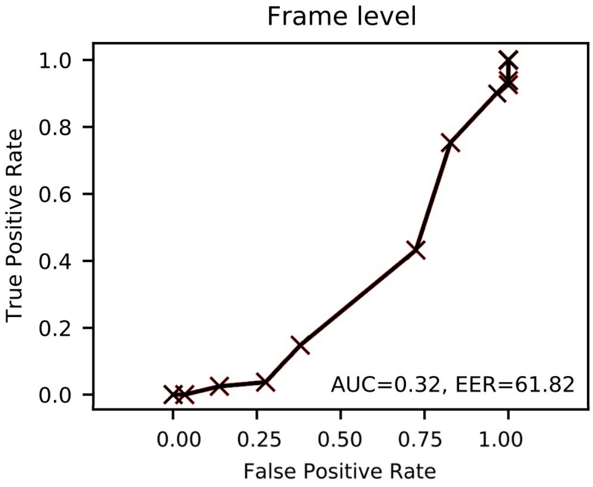


Figure 100: Ours$_{DOF(Brox)}$: frame-level ROC curve on Times Square View 2.

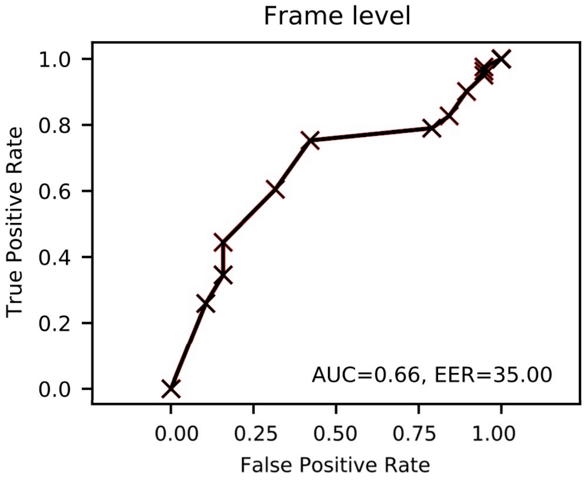


Figure 101: Ours$_{DOF(FlowNet)}$: frame-level ROC curve on Times Square View 2.

Compared to the detection results produced by applying the Liu et al. (2018b) and Vu et al. (2019) crowd anomaly detection methods on this scene, the results produced from the proposed framework indicate mediocre results. Section 6.1 details training and testing of these methods on the Times Square: View 2 scene. AUC and EER values given by evaluating Liu et al. (2018b) on this scene are 98.56 and 5.42 respectively. However, the AUC and EER values produced by training and testing Vu et al. (2019) on this scene are 66.02 and 34.54 respectively. While the detection results produced from applying Liu et al. (2018b) to this scene show better performance than the proposed method, our results, AUC of 70.37 and EER of 19.00, show a significant improvement in comparison to Vu et al. (2019).

6.4.3 Love Parade

The proposed framework is applied to the Love Parade scene from the AHDCrowd dataset for this experiment. The footage includes a high-density crowd with occurrences of over-crowding, crowd surges and a fight, the footage is shot from a wide-view angle. The proposed crowd anomaly detection method is trained and tested on this scene using the same four motion representation as the two previous experiments. The input given to the DAEs and its corresponding output are illustrated in Figure 102. As illustrated, the first column shows a sample image corrupted with noise and the corresponding reconstructed version. Additionally, the second column illustrates the dynamic image representation, FlowNet representation, dynamic optical flow (Brox) representation and dynamic optical flow (FlowNet) representation corresponding to the sample image corrupted with noise. Each motion representation has their corresponding reconstructed version illustrated alongside of it.

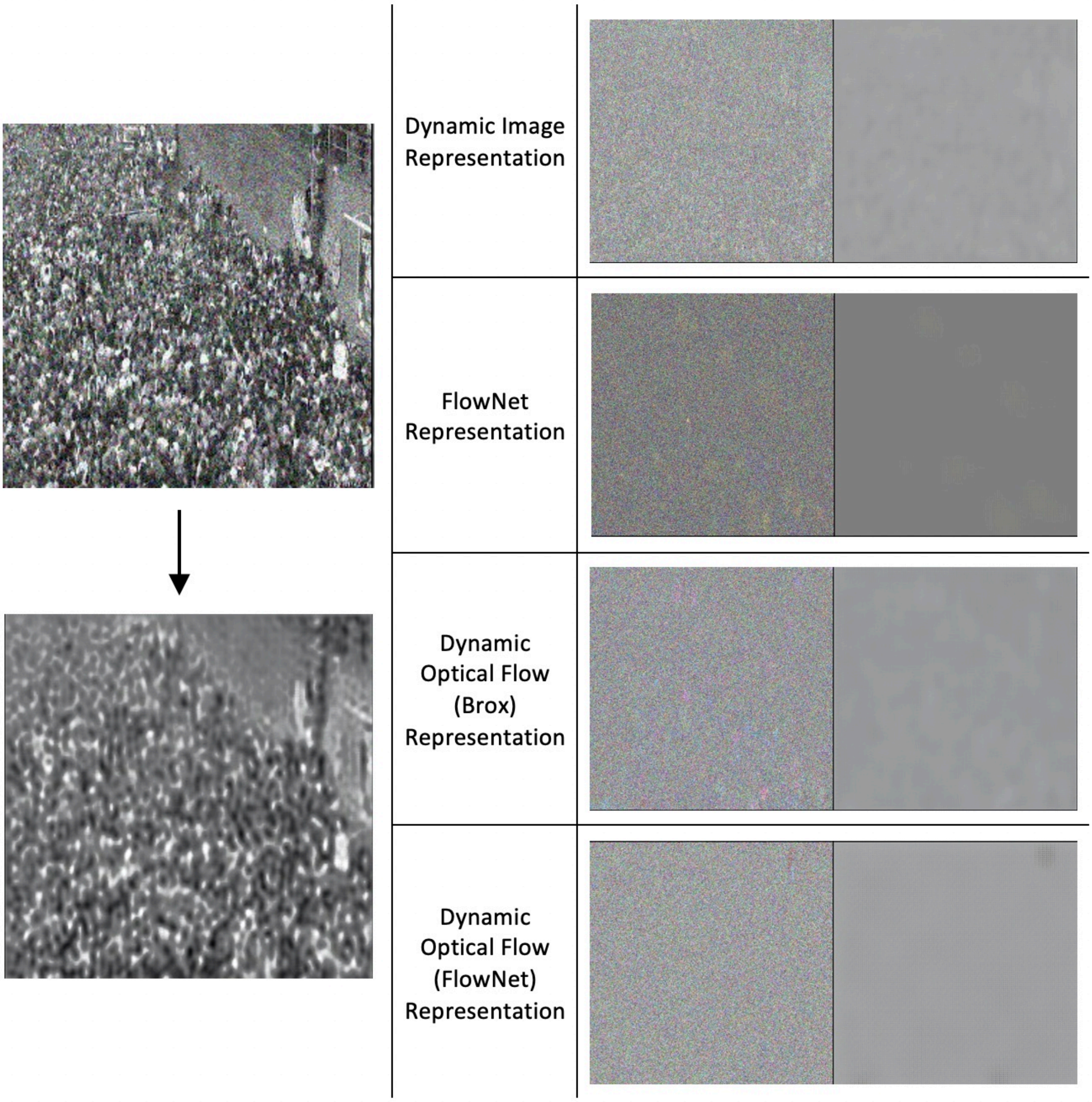

Dynamic Image Representation
FlowNet Representation
Dynamic Optical Flow (Brox) Representation
Dynamic Optical Flow (FlowNet) Representation

Figure 102: DAE reconstruction images of sample input frames (left) and the four corresponding motion representations (right) from the Love Parade scene.

The AUC, EER and ROC curve frame-level detection results of training and testing the proposed network on the Love Parade dataset are illustrated in Figures 103, 104, 105 and 106. These figures illustrate the results produced by using the motion representations: dynamic image, FlowNet, dynamic optical flow (Brox) and dynamic optical flow (FlowNet) respectively as input given to the proposed crowd anomaly detection framework. Table 13 shows the results produced using these motion representations. The best achieved AUC result is 96.15, this result is achieved by using dynamic optical flow (Brox) as the motion representation given to the proposed method. The AUC results produced using FlowNet also demonstrates competitive performance with an AUC value of 85.16. However, the remaining AUC results indicate that the performance of the other motion representatives are not of the same quality on this scene. Regarding the EER values, the best result, 9.00, is produced using FlowNet and dynamic optical flow (Brox) as motion representations. The remaining EER values given from using dynamic images and dynamic optical flow (FlowNet) as motion representations indicate a small decline in performance in comparison to the aforementioned EER results. Altogether, these results indicate that the most appropriate motion representation to be used with the proposed framework for the Love Parade scene is the dynamic optical flow (Brox) representation.

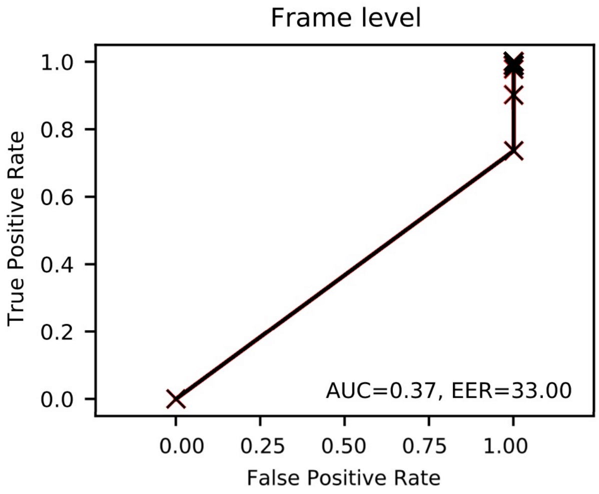


Figure 103: Ours$_{DI}$: frame-level ROC curve on Love Parade.

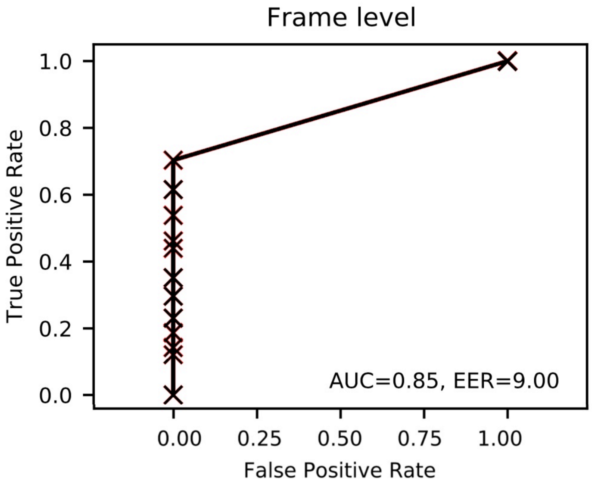


Figure 104: Ours$_{FlowNet}$: frame-level ROC curve on Love Parade.

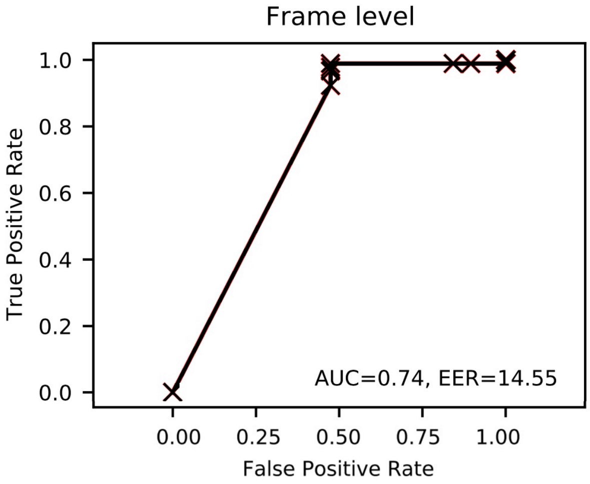


Figure 105: Ours$_{DOF(Brox)}$: frame-level ROC curve on Love Parade.

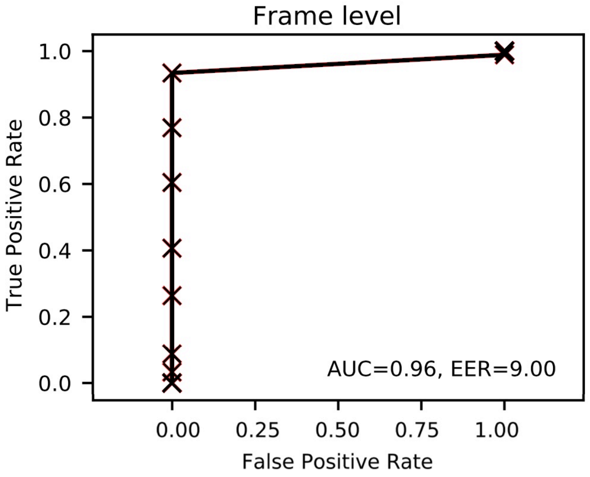


Figure 106: Ours$_{DOF(FlowNet)}$: frame-level ROC curve on Love Parade.

In comparison to the detection results produced by applying the crowd anomaly detection methods Liu et al. (2018b) and Vu et al. (2019) on this scene, the results produced by the proposed framework demonstrate a significant improvement in performance. Details of the training and testing process of these methods on the Love Parade scene are documented in Section 6.1. The AUC and EER values produced from the experiment of the Liu et al. (2018b) method on this scene are 55.24 and 47.14 respectively. Additionally, the AUC and EER values produced by evaluating the Vu et al. (2019) method on this scene are 88.08 and 16.36 respectively. Unlike the previous experiments, our results, AUC of 96.15 and EER of 9.00, indicate better performance than the detection results produced from applying Liu et al. (2018b) to this scene. Moreover, our results, demonstrate significant improvement in comparison to the detection results produced from applying Vu et al. (2019).

Below are ROC curves of each motion representation (Ours$_{DI}$, Ours$_{FlowNet}$, Ours$_{DOF(Brox)}$, Ours$_{DOF(FlowNet)}$) combined in a single image for easier viewing. Figures 107, 108, 109 are the frame-level ROC curves applied on the AHDCrowd Times Square: View 1, Times Square: View 2, and Love Parade datasets respectively.

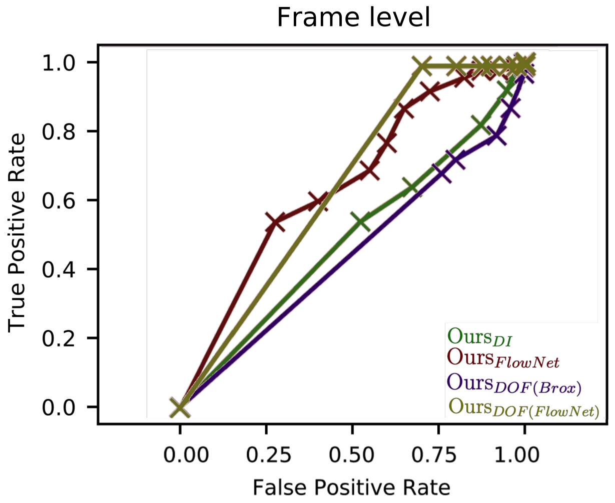


Figure 107: Frame-level ROCs on AHDCrowd Times Square: View 1 dataset.

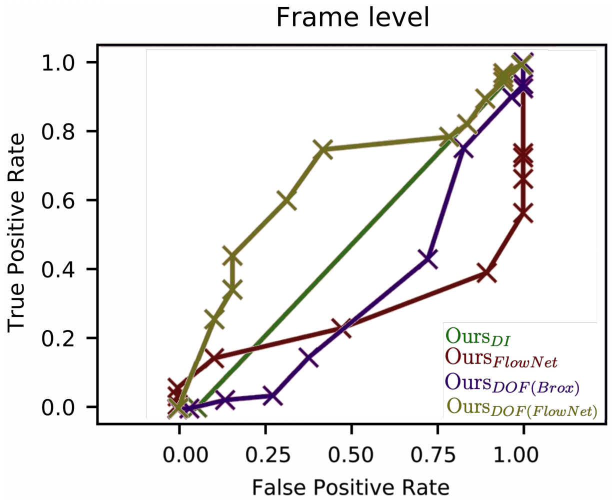


Figure 108: Frame-level ROCs on AHDCrowd Times Square: View 2 dataset.

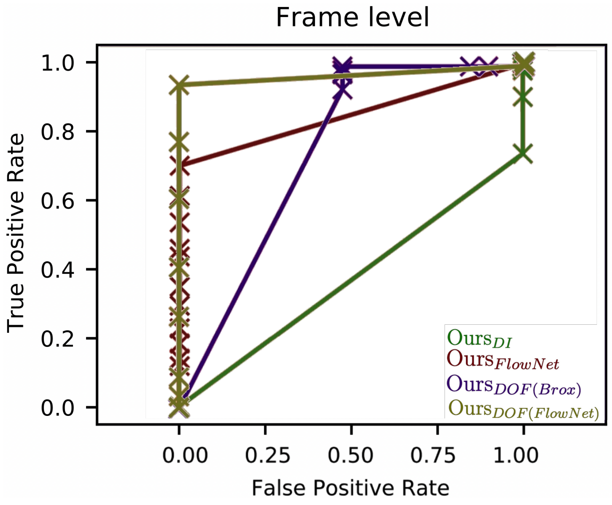


Figure 109: Frame-level ROCs on AHDCrowd Love Parade dataset.

6.5 Running times

Running times for crowd anomaly detection methods are not typically documented by their authors. Running times indicate the time taken to detect any anomalous scene withing a frame. The running times, frames per second (FPS), produced and documented by various methods have been reported in Table 14. These methods have been reviewed in Chapter 3. The running time for the crowd behaviour detection framework produced in this research (annotated as "Ours") is nearly 50 FPS.

Table 14: Running times of reviewed methods.

Name	FPS
Mahadevan et al. (2010)	0.4
Lu et al. (2013)	150
Li et al. (2014)	1.25
Sabokrou et al. (2015)	200
Chong and Tay (2017)	143
Sabokrou et al. (2018)	370
Liu et al. (2018b)	25
Ours	50

Various methods, Mahadevan et al. (2010), Li et al. (2014), and Liu et al. (2018b), have been able to achieve high running times. These methods have achieved running times that can be applied in the real world. CCTV footage captures 30 frames in a single second of video (30 FPS). In principle, any methods that can detect anomalous footage with running times less than 30 FPS can be applied in the real world. All experiments in this thesis are carried out on Google Colab in 2019. The running times documented in Table 14 are collected from the authors papers as well as Ramachandra et al. (2020) and Xu et al. (2019). These running times are a rough indicators of the performance of these methods, some methods prioritise higher performance evaluation metrics such as AUC, EER and ROC curves as opposed to the running times. Moreover, the running times collected are not up to date because these results are based on the publication dates of these methods and the authors carry out their experiments on different machines.

6.6 Conclusion

This chapter has described the experiments that were carried out to evaluate the contributions proposed in this research. A novel high-density crowd dataset containing normal and abnormal crowd behaviour was created as one of these contributions. This dataset, to the best of this researcher's knowledge, is the only dataset with footage of typical behaviour high-density crowds and also includes annotated occurrences of anomalous behaviour. The dataset was used to train and test state-of-the-art crowd anomaly detection methods. The experimentation presented in this chapter has tested three crowd abnormality detection methods: Spatiotemporal Autoencoder (Chong and Tay, 2017), Future Frame Prediction (Liu et al., 2018b), and Anomaly Detection Using Multilevel Representations (Vu et al., 2019). They were tested to evaluate their performance (Regularity score, AUC, and EER) with this dataset. The results produced by these methods when modelled on a highly dense crowd suggest that the transition from low-medium density crowd abnormality detection into high-density crowd abnormality detection has weakened their performance. Moreover, these results have demonstrated the necessity for crowd anomaly detection methods to take into consideration high-density crowds in training/testing.

The remaining contributions are the development of a CGAN architecture combined with Dynamic Images (Bilen et al., 2016) for crowd behaviour anomaly detection, and using CGANs to distinguish between normal and abnormal behaviour within high-density crowds. These contributions have been evaluated by testing the proposed crowd anomaly detection framework proposed in this research. The framework combines Dynamic Images and image-to-image translation using CGANs as a novel approach for the detection of anomalous behaviour within a crowd. The framework was evaluated using four different motion representations; dynamic image representation, FlowNet representation, dynamic optical flow representation (computed using Brox) and dynamic optical flow representation (computed using FlowNet). The achieved results have demonstrated the merits and faults of utilising each motion representation into the proposed framework. The overall results have shown the advantages of utilising dynamic image representations to calculate the temporal development of a video as an alternative to optical flow. Particularly when tested on high-density crowds. Moreover, the detection results using FlowNet instead of Brox, as the motion representation, have shown the merits of FlowNet integrated into the proposed framework. Specifically when tested on low to medium-density crowds.

Finally, additional experimentation of the proposed crowd anomaly detection method was applied by utilising the AHDCrowd dataset. Due to the scarcity of public high-density anomalous crowd datasets, three scenes from the AHDCrowd dataset were used to train and test the proposed architecture. The framework was evaluated using four different motion representations: dynamic image representation, FlowNet representation, dynamic optical flow representation (computed using Brox) and dynamic optical flow representation (computed using FlowNet). The results indicated that the use of dynamic images and dynamic optical flow (FlowNet) as motion representations given to the proposed method do not perform well when applied on high-density crowds. However, the use of FlowNet optical flow has achieved better results and dynamic optical flow (Brox) has outperformed (AUC and EER) all the other motion representations.

7 Discussions and Conclusions

7.1 Discussion

Crowd behaviour analysis and anomaly detection are key to effective intelligent vision systems. Anomaly detection can be achieved by the detecting atypical patterns or detecting sudden changes in crowd flow/behaviour. This thesis focuses on anomaly detection in high-density crowds and has proposed a novel crowd anomaly detection framework "Dynamic Image Crowd Representations for Improved Anomaly Detection using Generative Adversarial Networks". The proposed framework combines Dynamic Images and image-to-image translation using CGANs as a novel approach for the detection of anomalous crowd behaviour. As an alternative to optical flow extraction using Brox et al. (2004), commonly used in state-of-the-art methods, this proposed framework utilises dynamic optical flow representations to extract the temporal development for a set of images. The extracted temporal features are used as the motion representation incorporated into the proposed anomaly detection framework.

The experiments that have been carried out in this research were used to evaluate the efficiency of the proposed framework for detecting anomalies in high-density crowds. A new high-density crowd dataset containing crowd behaviour anomalies (AHDCrowd) was created as no such datasets currently exist. Initially, the dataset was used to train and test state-of-the-art crowd anomaly detection methods which included Spatiotemporal Autoencoder Chong and Tay (2017), Future Frame Prediction Liu et al. (2018b), and Anomaly Detection Using Multilevel Representations Vu et al. (2019). The experiments used three scenes from the AHDCrowd dataset: Times Square View 1, Times Square View 2 and Love Parade scenes. These scenes included footage of anomalous crowd behaviours where crowds disperse quickly and frantically or where fights occur. The evaluation results produced by applying the method proposed by Chong and Tay (2017) indicate that the transition from low-medium density crowd anomaly detection into high-density crowd abnormality detection has weakened its performance. The plotted regularity scores demonstrate normal behaviour occurrences that do not conform with the dataset's ground-truth data. The results generated from the application of Liu et al. (2018b) and Vu et al. (2019) on the AHDCrowd dataset have demonstrated better detection results. However, compared to their results on the benchmark low to medium-density data sets, the detection performance has weakened. These results established the necessity for crowd anomaly detection methods to consider high-density crowds in the training and testing processes.

The novel crowd anomaly detection framework was evaluated by training and testing it on benchmark datasets, the results were compared to state-of-the-art crowd anomaly detection methods. The framework is evaluated using four different motion representations: dynamic image representation, FlowNet representation, dynamic optical flow representation (computed using Brox) and dynamic optical flow representation (computed using FlowNet). Dynamic image representations demonstrated low performance compared to state-of-the-art; using dynamic image extraction on the raw input data reduced image quality generated by the CGAN. FlowNet, a more novel approach for optical flow computation, is used as the motion representation incorporated into the subsequent experiment's framework. The detection results produced from this experiment has outperformed the existing state-of-the-art on UCSD Ped-2 and Avenue datasets and produced comparable results on the UCSD Ped-1 dataset. Pixel-level detection particularly achieved excellent results, demonstrating the capabilities of this experiment in localising the detected anomalies.

As previously surveyed, dynamic optical flow outperforms optical flow and dynamic image representations in the field of action recognition. Consequently, dynamic optical flow representations are incorporated into the proposed framework for the next experiment. The dynamic images are extracted from pre-computed optical flow maps using Brox optical flow method. Although pixel-level detection results on UCSD Ped-1 show a decline in performance, the results on the UCSD Ped-2 have demonstrated results on par with other methods. Results on frame-level detection have shown similar performance results on the UCSD Ped-1 dataset, but the UCSD Ped-2 and Avenue dataset results are either on par or higher than the other methods. Finally, dynamic optical flow representations using FlowNet for optical flow extraction were incorporated into the last experiment. Similar to the previous experiment, pixel and frame-level detection results on the UCSD Ped-1 dataset demonstrate comparable performance. Moreover, frame-level detection on the UCSD Ped-2 and Avenue datasets has outperformed the state-of-the-art regarding AUC and EER values. These results demonstrate the advantages and disadvantages of incorporating each motion representation into the proposed framework. Overall the results have shown the advantages of utilising dynamic image representations to calculate the temporal development of a video as an alternative to optical flow. Incorporating dynamic optical flow (FlowNet) representations have improved detection results on frame-level while incorporating just optical flow extracted from FlowNet has improved pixel-level detection results.

The final experiments evaluate the proposed framework's performance when applied to a high-density crowd. Three scenes from the AHDCrowd dataset are used to train and test the proposed architecture. Similar to the previous experiments, the method was evaluated using four different motion representations to calculate an input video's temporal development. The motion representation used is dynamic image representation, FlowNet representation, dynamic optical flow representation (computed using Brox) and dynamic optical flow representation (computed using FlowNet). The detection results produced by applying the proposed crowd anomaly detection method on the Times Square: View 1 scene show that using dynamic optical flow (Brox) as the motion representation outperforms the remaining motion representations. Similar performance results were achieved using the two remaining scenes; Times Square: View 1 and Love Parade. Utilising dynamic optical flow (Brox) as the motion representation input to the proposed method achieves the best detection performance on a high-density crowd. Moreover, compared to the detection results produced from the application of several state-of-the-art crowd anomaly detection methods on these scenes from the dataset, the proposed framework is more competitive and has outperformed the other methods.

The novel crowd anomaly detection method proposed in this research has not been trained and tested on a more general-purpose setting. Similar to this research, state-of-the-art crowd anomaly detection methods have used benchmark datasets to train and test their methods. Our research and the SOA in crowd anomaly detection have not focused on a general-purpose setting to be able to compare results with other methods. Some of the general purpose setting applications can be training a method on different scenes from different cameras. Alternately, the training process remains the same but the testing process could be applied on a new location different to the training data. Training and testing the method proposed in this research on a more general-purpose setting is part of the future work proposed in Section 7.3.

7.2 Conclusion

As this research's focus is to detect anomalous behaviour within a crowd utilising computer vision and machine learning methods, a comprehensive overview of crowd analysis and crowd behaviour analysis was applied. Fields such as crowd counting, density estimation, crowd tracking, person re-identification, motion representation and anomaly detection were investigated. Improvements can be applied to algorithms regarding crowd counting and density estimation. Some of the major issues found were severe occlusion handling, adaptability towards static and dynamic movements of people or objects, environmental changes such as weather and illumination variations. Moreover, progress is still to be achieved regarding tracking and re-identification; biometric data has not been effectively incorporated to tracking and re-identification algorithms. Representational models of the links between low-level features and high-level features have not been sufficiently integrated with these algorithms. Additionally, the limitations in crowd anomaly detection methods became apparent; the accuracy results presented by previous work were not satisfactory enough to be applied to the real-world environment. The contributions of this research indicated using bold lettering, have been met as follows.

The development of a CGAN architecture combined with Dynamic Images (Bilen et al., 2016) provides a novel approach for crowd behaviour anomaly detection.

State-of-the-art methods are investigated, and generative adversarial networks (GANs), more specifically, image-to-image translation using Conditional GANs for anomaly detection, displayed their potential and could benefit from a more thorough investigation. The application of GANs in crowd anomaly detection has proved to achieve higher performance than earlier methods. Consequently, this architecture was chosen as the base of the framework proposed in this research. A novel approach to crowd anomaly detection utilising Dynamic Images was investigated. Dynamic optical flow representations were used as motion representations and incorporated to the proposed framework. The framework was applied and evaluated on benchmark crowd anomaly detection datasets to evaluate its performance compared to state-of-the-art methods in this field. The evaluation results produced are marginally higher than those produced by the state-of-the-art.

This research also addresses a substantial gap concerning the evaluation of crowd anomaly detection methods with highly dense crowds. Benchmark datasets include footage of low to medium-density crowds, but datasets including high-density crowds with anomalous behaviour are not published. Therefore, as another contribution to this research, **A labelled high-density crowd dataset containing normal and abnormal (footage with anomalous behaviour) was created for this purpose. The dataset has been applied to anomaly detection algorithms and has been made public to other researchers.**

This dataset was created by collecting, processing, and labelling footage of environments containing highly dense crowds and occurrences of anomalies. State-of-the-art crowd anomaly detection methods were trained and tested on this dataset to evaluate the difference in performance when transitioning from low to medium-density crowds into high-density crowds.

Generative modelling for anomaly detection in high-density crowds. Conditional Generative Adversarial Networks (CGANs) produces data to a discriminative function to distinguish between normal and abnormal behaviour within medium to high-density crowds.

Additional experiments were applied using the proposed crowd anomaly detection method on the Abnormal High-Density Crowd dataset to evaluate the method's performance in anomaly detection. The crowd detection results demonstrate that applying the proposed method performs well when applied on high-density crowds.

The contributions of this research have been achieved and the applied experiments evaluate the effectiveness of utilising dynamic image representations for anomaly detection within low-medium and high-density crowds. Currently, this research can detect and localise anomalies in footage allowing the recognition of when and where an anomaly occurs. This is beneficial to understand the start, end, and location of anomalies for further investigation. As the processing power increases and running times decrease, this research can be applied in the real-world to help in the prevention of chaotic events (abnormal behaviour).

7.3 Future Work

Further advancement in the area of crowd anomaly detection must consider high-density crowds in training and testing crowd anomaly detection models to detect anomalous behaviour in highly dense environments effectively. At the time of writing this thesis, high-density crowds have not been given adequate attention. Benchmark datasets used to train and test novel crowd anomaly detection methods consider low and medium density crowds. The Abnormal High-Density Crowd dataset (AHDCrowd) created in this research will help future researchers create anomaly detection methods applicable to high-density crowds. Another limitation in crowd anomaly detection is that benchmark datasets do not include different anomalous behaviour types and the scenes are enacted, limiting accurate evaluation of crowd anomaly detection methods. Future researchers can increase the types of anomalies used to train and test crowd anomaly detection methods by collecting footage of various real-world anomaly occurrences. The footage can be structured and labelled using the same approach used to create the Abnormal High-Density Crowd dataset in this research.

Other advancements to this thesis are the continuation of training and testing through an ablation study of the proposed framework. The ablation study would be applied to asses the performance of the method when certain components of the framework are removed. This would help in the understanding of the contribution of the removed component on the framework. Moreover, due to COVID-19, access to a High Performance Computing (HPC) system was not feasible while conducting the experiments documented in this research. So to conduct an ablation study swiftly and effectively the use of an HPC system would be recommended. Also, with the use of an HPC system a more general-purpose application of this framework could be established. The proposed method could be trained on scenes from different cameras and find how it will effect the anomaly detection results. Alternately, another experiment could be applied where the training process remains the same but the testing process could be applied on a location different to the training data. These are some experiments that could be conducted to understand the behaviour of the proposed method on a more general-purpose setting.

Additionally, running times should be decreased to pursue real-time crowd anomaly detection with comparable AUC and EER results as state-of-the-art. Currently, running times of novel anomaly detection methods do not meet the real-time application requirements. Furthermore, dynamic image representations should be used for temporal development extraction in other anomaly crowd detection methods instead of optical flow difference. In the field of action recognition, the use of dynamic image representations have outperformed the use of optical flow extractions. Additionally, this research demonstrates the effectiveness of replacing optical flow extractions, the conventional temporal development extraction method, with the more novel dynamic image representations in the field of crowd anomaly detection.

References

Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D., 2008. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 555–560. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-39549091484, doi:10.1109/TPAMI.2007.70825.

Advanced Video and Signal based Surveillance, 2007. i-lids bag and vehicle detection dataset. Available at http://www.eecs.qmul.ac.uk/{~}andrea/avss2007_d.html. [Accessed 21 November 2018].

Afsar, P., Cortez, P., Santos, H., 2015. Automatic visual detection of human behavior: A review from 2000 to 2014. Expert Systems with Applications 42, 6935–6956. URL: http://www.sciencedirect.com/science/article/pii/S0957417415003516, doi:https://www.doi.org/10.1016/j.eswa.2015.05.023.

Ali, S., Shah, M., 2007. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–6. doi:10.1109/CVPR.2007.382977.

Ali, S., Shah, M., 2008. Floor fields for tracking in high density crowd scenes, in: European conference on Computer Vision, Springer. pp. 1–14.

Andrade, E.L., Blunsden, S., Fisher, R.B., 2006. Modelling crowd scenes for event detection, in: 18th International Conference on Pattern Recognition (ICPR'06), pp. 175–178. doi:10.1109/ICPR.2006.806.

Arjovsky, M., Bottou, L., 2017. Towards principled methods for training generative adversarial networks. arXiv preprint arXiv:1701.04862 .

Assari, S.M., Idrees, H., Shah, M., 2016. Human re-identification in crowd videos using personal, social and environmental constraints, in: European Conference on Computer Vision, Springer. pp. 119–136.

Baltieri, D., Vezzani, R., Cucchiara, R., Utasi, Á., Benedek, C., Szirányi, T., 2011. Multi-view people surveillance using 3D information, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1817–1824. doi:10.1109/ICCVW.2011.6130469.

Bazzani, L., Cristani, M., Murino, V., 2013. Symmetry-driven accumulation of local features for human characterization and re-identification. Computer Vision and Image Understanding 117, 130–144. URL: http://www.sciencedirect.com/

science/article/pii/S1077314212001464. doi:https://www.doi.org/10.1016/j.cviu.2012.10.008.

Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V., 2010. Multiple-shot person re-identification by hpe signature, in: 2010 20th International Conference on Pattern Recognition, pp. 1413–1416. doi:10.1109/ICPR.2010.349.

Bera, A., Manocha, D., 2018. Interactive surveillance technologies for dense crowds. arXiv preprint arXiv:1810.03965 .

Berclaz, J., Fleuret, F., Turetken, E., Fua, P., 2011. Multiple Object Tracking using K-Shortest Paths Optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence , 1806–1819.

van den Berg, J., Guy, S.J., Lin, M., Manocha, D., 2011. Reciprocal n-body collision avoidance, in: Pradalier, C., Siegwart, R., Hirzinger, G. (Eds.), Robotics Research, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 3–19.

Bertini, M., Del Bimbo, A., Seidenari, L., 2012. Scene and crowd behaviour analysis with local space-time descriptors, in: 2012 5th International Symposium on Communications, Control and Signal Processing, IEEE. pp. 1–6.

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., 2017. Action recognition with dynamic image networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 40, 2799–2813.

Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S., 2016. Dynamic image networks for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3034–3042.

Blunsden, S., Fisher, R., 2010. The behave video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA 4, 4.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D., 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 3722–3731.

Brostow, G.J., Cipolla, R., 2006. Unsupervised bayesian detection of independent motion in crowds, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 594–601. doi:10.1109/CVPR.2006.320.

Brox, T., Bruhn, A., Papenberg, N., Weickert, J., 2004. High accuracy optical flow estimation based on a theory for warping, in: European conference on Computer Vision, Springer. pp. 25–36.

Candamo, J., Shreve, M., Goldgof, D.B., Sapper, D.B., Kasturi, R., 2010. Understanding transit scenes: A survey on human behavior-recognition algorithms. IEEE Transactions on Intelligent Transportation Systems 11, 206–224. doi:10.1109/TITS.2009.2030963.

Cao, T., Wu, X., Guo, J., Yu, S., Xu, Y., 2009. Abnormal crowd motion analysis, in: 2009 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1709–1714. doi:10.1109/ROBIO.2009.5420408.

CAVIAR, 2003. Context aware vision using image-based active recognition. Available at http://www.homepages.inf.ed.ac.uk/rbf/CAVIAR/. [Accessed 21 November 2017].

Chan, A.B., Liang, Z.J., Vasconcelos, N., 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–7. doi:10.1109/CVPR.2008.4587569.

Chan, A.B., Vasconcelos, N., 2009. Bayesian poisson regression for crowd counting, in: 2009 IEEE 12th International Conference on Computer Vision, pp. 545–551. doi:10.1109/ICCV.2009.5459191.

Chapelle, O., Haffner, P., Vapnik, V.N., 1999. Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks 10, 1055–1064. doi:10.1109/72.788646.

Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A., 2013. A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding 117, 633–659. URL: http://www.sciencedirect.com/science/article/pii/S1077314213000295, doi:https://www.doi.org/10.1016/j.cviu.2013.01.013.

Chen, D., Huang, P., 2013. Visual-based human crowds behavior analysis based on graph modeling and matching. IEEE Sensors Journal 13, 2129–2138. doi:10.1109/JSEN.2013.2245889.

Chen, K., Loy, C.C., Gong, S., Xiang, T., 2012. Feature mining for localised crowd counting, in: Proceedings of the British Machine Vision Conference, BMVA Press. pp. 21.1–21.11. doi:http://dx.doi.org/10.5244/C.26.21.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems, pp. 2172–2180.

Cheriyadat, A.M., Radke, R.J., 2008. Detecting dominant motions in dense crowds. IEEE Journal of Selected Topics in Signal Processing 2, 568–581. URL: http://www.ecse.rpi.edu/~rjradke/jstsp/, doi:10.1109/JSTSP.2008.2001306.

Cho, S., Chow, T.W.S., 1999. A fast neural learning vision system for crowd estimation at underground stations platform. Neural Processing Letters 10, 111–120. URL: https://www.doi.org/10.1023/A:1018781301409, doi:10.1023/A:1018781301409.

Cho, S., Chow, T.W.S., Leung, C., 1999. A neural-based crowd estimation by hybrid global learning algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29, 535–541. doi:10.1109/3477.775269.

Chong, Y.S., Tay, Y.H., 2017. Abnormal event detection in videos using spatiotemporal autoencoder, in: International Symposium on Neural Networks, Springer. pp. 189–196.

Cristani, M., Raghavendra, R., Bue, A.D., Murino, V., 2013. Human behavior analysis in video surveillance: A social signal processing perspective. Neurocomputing 100, 86–97. URL: http://www.sciencedirect.com/science/article/pii/S0925231212003141, doi:https://www.doi.org/10.1016/j.neucom.2011.12.038. special issue: Behaviours in video.

Cui, X., Liu, Q., Gao, M., Metaxas, D.N., 2011. Abnormal detection using interaction energy potentials, in: CVPR 2011, pp. 3161–3167. doi:10.1109/CVPR.2011.5995558.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: 2005 IEEE computer society conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 886–893.

Dalal, N., Triggs, B., Schmid, C., 2006. Human detection using oriented histograms of flow and appearance, in: European conference on Computer Vision, Springer. pp. 428–441.

Dee, H.M., Caplier, A., 2010. Crowd behaviour analysis using histograms of motion direction, in: 2010 IEEE International Conference on Image Processing, pp. 1545–1548. doi:10.1109/ICIP.2010.5653573.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 248–255.

Denton, E., Gross, S., Fergus, R., 2016. Semi-supervised learning with context-conditional generative adversarial networks. arXiv preprint arXiv:1611.06430 .

Dickinson, P., Gerling, K., Hicks, K., Murray, J., Shearer, J., Greenwood, J., 2019. Virtual reality crowd simulation: effects of agent density on user experience and behaviour. Virtual Reality 23, 19–32.

Donahue, J., Krähenbühl, P., Darrell, T., 2016. Adversarial feature learning. arXiv preprint arXiv:1605.09782 .

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T., 2015. Flownet: Learning optical flow with convolutional networks, in: Proceedings of the IEEE international conference on Computer Vision, pp. 2758–2766.

Dupont, C., Tobias, L., Luvison, B., 2017. Crowd-11: A dataset for fine grained crowd behaviour analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 9–16.

Dupre, R., Argyriou, V., 2019. A human and group behavior simulation evaluation framework utilizing composition and video analysis. Computer Animation and Virtual Worlds 30, e1844.

El-Etriby, S., Miraoui, M., Elmezain, M., 2017. Crowd behavior analysis using discriminative models in video sequences. Journal of Computational and Theoretical Nanoscience 14, 6–2714. URL: http://www.ingentaconnect.com/content/asp/jctn/2017/00000014/00000006/art00016, doi:doi:10.1166/jctn.2017.6559.

Ess, A., Leibe, B., Gool, L.V., 2007. Depth and appearance for mobile scene analysis, in: 2007 IEEE 11th International Conference on Computer Vision, pp. 1–8. doi:10.1109/ICCV.2007.4409092.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press. p. 226–231.

Fan, Y., Wen, G., Li, D., Qiu, S., Levine, M.D., Xiao, F., 2020. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. Computer Vision and Image Understanding 195, 102920. doi:https://doi.org/10.1016/j.cviu.2020.102920.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010. Person re-identification by symmetry-driven accumulation of local features, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2360–2367. doi:10.1109/CVPR.2010.5539926.

Farnebäck, G., 2003. Two-frame motion estimation based on polynomial expansion, in: Bigun, J., Gustavsson, T. (Eds.), Image Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 363–370.

Ferryman, J., Shahrokni, A., 2009. Pets2009: Dataset and challenge, in: 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 1–6. doi:10.1109/PETS-WINTER.2009.5399556.

Ferryman, J., Tweed, D., 2007. An overview of the pets 2007 dataset, in: Proceeding Tenth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS, p. 49–53.

Figueiredo, M., Jain, A., 2002. Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis & Machine Intelligence 24, 381–396. URL: doi.ieeecomputersociety.org/10.1109/34.990138, doi:10.1109/34.990138.

Fradi, H., Luvison, B., Pham, Q.C., 2017. Crowd behavior analysis using local mid-level visual descriptors. IEEE Transactions on Circuits and Systems for Video Technology 27, 589–602. doi:10.1109/TCSVT.2016.2615443.

Gandhi, T., Trivedi, M.M., 2007. Person tracking and reidentification: Introducing panoramic appearance map (pam) for feature representation. Machine Vision and Applications 18, 207–220. URL: https://www.doi.org/10.1007/s00138-006-0063-x, doi:10.1007/s00138-006-0063-x.

Gao, G., Gao, J., Liu, Q., Wang, Q., Wang, Y., 2020. Cnn-based density estimation and crowd counting: A survey. ArXiv abs/2003.12783.

Ge, W., Collins, R.T., 2009. Marked point processes for crowd counting, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2913–2920. doi:10.1109/CVPR.2009.5206621.

Georgescu, M.I., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M., 2020. Anomaly detection in video via self-supervised and multi-task learning. arXiv preprint arXiv:2011.07491 .

Goodfellow, I., 2016. Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160 .

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems, pp. 2672–2680.

Goodnight, J., 2018. Introduction to SEMMA. Available at https://www.documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en. [Accessed 21 November 2018].

Grant, J.M., Flynn, P.J., 2017. Crowd scene understanding from video: A survey. ACM Trans. Multimedia Comput. Commun. Appl. 13, 19:1–19:23. URL: http://www.doi.acm.org/10.1145/3052930, doi:10.1145/3052930.

Graves, A., Jaitly, N., Mohamed, A., 2013. Hybrid speech recognition with deep bidirectional lstm, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278. doi:10.1109/ASRU.2013.6707742.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein gans, in: Advances in Neural Information Processing Systems, pp. 5767–5777.

Haghani, M., Sarvi, M., 2018. Crowd behaviour and motion: Empirical methods. Transportation research part B: methodological 107, 253–294.

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A.K., Davis, L.S., 2016. Learning temporal regularity in video sequences, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 733–742.

Hassner, T., Itcher, Y., Kliper-Gross, O., 2012. Violent flows: Real-time detection of violent crowd behavior, in: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1–6. doi:10.1109/CVPRW.2012.6239348.

Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: A survey. Image and Vision Computing 60, 4–21. doi:https://doi.org/10.1016/j.imavis.2017.01.010.

Hergott, M., 2019. A leap into the future: Generative adversarial networks. Available at https://www.medium.com/datadriveninvestor/a-leap-into-the-future-generative-adversarial-networks-96a780ed8ee6. [Accessed 21 November 2019].

Hjelm, R.D., Jacob, A.P., Che, T., Trischler, A., Cho, K., Bengio, Y., 2017. Boundary-seeking generative adversarial networks. arXiv preprint arXiv:1702.08431 .

Horn, B.K., Schunck, B.G., 1981. Determining optical flow, in: Techniques and Applications of Image Understanding, International Society for Optics and Photonics. pp. 319–331.

Idrees, H., Saleemi, I., Seibert, C., Shah, M., 2013. Multi-source multi-scale counting in extremely dense crowd images, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 2547–2554.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR abs/1502.03167. URL: http://www.arxiv.org/abs/1502.03167.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134.

Javed, O., Shafique, K., Rasheed, Z., Shah, M., 2008. Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding 109, 146–162. URL: http://www.sciencedirect.com/science/article/pii/S1077314207000100, doi:https://www.doi.org/10.1016/j.cviu.2007.01.003.

Ji, S., Xu, W., Yang, M., Yu, K., 2012. 3d convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 35, 221–231.

Ji, X., Li, B., Zhu, Y., 2020. Tam-net: Temporal enhanced appearance-to-motion generative network for video anomaly detection, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–8.

Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D., Shao, L., 2019. Crowd counting and density estimation by trellis encoder-decoder networks, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6133–6142.

Joshi, K.V., Patel, N.M., Shah, M.B., 2019. A review on crowd analysis techniques. Journal of Artificial Intelligence Research & Advances 6, 119–126.

Junior, J.C.S.J., Musse, S.R., Jung, C.R., 2010. Crowd analysis using computer vision techniques. IEEE Signal Processing Magazine 27, 66–77. doi:10.1109/MSP.2010.937394.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME Journal of Basic Engineering , 35–45.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 .

Kasturi, R., Ekambaram, R., 2014. Person reidentification and recognition in video, in: Iberoamerican Congress on Pattern Recognition, Springer. pp. 280–293.

Ko, T., 2008. A survey on behavior analysis in video surveillance for homeland security applications, in: 2008 37th IEEE Applied Imagery Pattern Recognition Workshop, pp. 1–8. doi:10.1109/AIPR.2008.4906450.

Kok, V.J., Lim, M.K., Chan, C.S., 2016. Crowd behavior analysis: A review where physics meets biology. Neurocomputing 177, 342–362. URL: http://www.sciencedirect.com/science/article/pii/S0925231215017403, doi:https://www.doi.org/10.1016/j.neucom.2015.11.021.

Kong, D., Gray, D., Tao, H., 2006. A viewpoint invariant approach for crowd counting, in: 18th International Conference on Pattern Recognition (ICPR'06), pp. 1187–1190. doi:10.1109/ICPR.2006.197.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, pp. 1097–1105.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., 2011. Hmdb: a large video database for human motion recognition, in: 2011 International Conference on Computer Vision, IEEE. pp. 2556–2563.

Lamba, S., Nain, N., 2017. Crowd monitoring and classification: a survey, in: Advances in Computer and Computational Sciences. Springer, pp. 21–31.

Lavi, B., Serj, M.F., Ullah, I., 2018. Survey on deep learning techniques for person re-identification task. CoRR abs/1807.05284. URL: http://arxiv.org/abs/1807.05284.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., Yan, S., 2017. Perceptual generative adversarial networks for small object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1222–1230.

Li, T., Chang, H., Wang, M., Ni, B., Hong, R., Yan, S., 2015. Crowded scene analysis: A survey. IEEE Transactions on Circuits and Systems for Video Technology 25, 367–386. doi:10.1109/TCSVT.2014.2358029.

Li, W., Mahadevan, V., Vasconcelos, N., 2014. Anomaly detection and localization in crowded scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 18–32. doi:10.1109/TPAMI.2013.111.

Liang, R., Zhu, Y., Wang, H., 2014. Counting crowd flow based on feature points. Neurocomputing 133, 377–384. URL: http://www.sciencedirect.com/science/article/pii/S0925231214000630, doi:https://www.doi.org/10.1016/j.neucom.2013.12.040.

Lim, M.K., Kok, V.J., Loy, C.C., Chan, C.S., 2014. Crowd saliency detection via global similarity structure, in: 2014 22nd International Conference on Pattern Recognition, IEEE. pp. 3957–3962.

Lin, S., Chen, J., Chao, H., 2001. Estimation of number of people in crowded scenes using perspective transformation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 31, 645–654. doi:10.1109/3468.983420.

Liu, J., Gao, C., Meng, D., Hauptmann, A.G., 2018a. Decidenet: Counting varying density crowds through attention guided detection and density estimation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197–5206.

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

Liu, X., van de Weijer, J., Bagdanov, A.D., 2018c. Leveraging unlabeled data for crowd counting by learning to rank, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7661–7669.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110. URL: https://www.doi.org/10.1023/B:VISI.0000029664.99615.94, doi:10.1023/B:VISI.0000029664.99615.94.

Lu, C., Shi, J., Jia, J., 2013. Abnormal event detection at 150 fps in matlab, in: 2013 IEEE International Conference on Computer Vision, pp. 2720–2727. doi:10.1109/ICCV.2013.338.

Luc, P., Couprie, C., Chintala, S., Verbeek, J., 2016. Semantic segmentation using adversarial networks. CoRR abs/1611.08408. URL: http://www.arxiv.org/abs/1611.08408.

Lucas, B.D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. Proceedings DARPA image Understanding Workshop , 121–130.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proceedings of the 30th International Conference on Machine Learning, p. 3.

Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N., 2010. Anomaly detection in crowded scenes, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1975–1981. doi:10.1109/CVPR.2010.5539872.

Mahmoud, S., 2019. Abnormal high-density crowd dataset. Available at https://www.kaggle.com/angelchi56/abnormal-highdensity-crowds. [Accessed 01 January 2020].

Mahmoud, S., 2020. Dynamic image crowd representations for improved anomaly detection using generative adversarial networks. Available at https://drive.google.com/drive/folders/1eFeC3cRVCF6GUjgc_PB_40G83vsZ4_yW?usp=sharing. [Accessed 01 January 2021].

Mahmoud, S., Arafa, Y., 2020. Abnormal high-density crowd dataset, in: 2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA), pp. 57–65. doi:10.1109/MCNA50957.2020.9264277.

Majumder, A., Babu, R.V., Chakraborty, A., 2018. Anomaly from motion: Unsupervised extraction of visual irregularity via motion prediction, in: Rameshan, R., Arora, C., Dutta Roy, S. (Eds.), Computer Vision, Pattern Recognition, Image Processing, and Graphics, Springer Singapore, Singapore. pp. 66–77.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B., 2015. Adversarial autoencoders. arXiv preprint arXiv:1511.05644 .

Martin A., S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, pp. 214–223.

Matkovic, F., Marčetic, D., Ribaric, S., 2019. Abnormal crowd behaviour recognition in surveillance videos, in: 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), IEEE. pp. 428–435.

Mazzon, R., Tahir, S.F., Cavallaro, A., 2012. Person re-identification in crowd. Pattern Recognition Letters 33, 1828—1837. doi:10.1016/j.patrec.2012.02.014.

Mehran, R., Oyama, A., Shah, M., 2009. Abnormal crowd behavior detection using social force model, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 935–942. doi:10.1109/CVPR.2009.5206641.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. CoRR abs/1411.1784. URL: http://www.arxiv.org/abs/1411.1784.

Musse, S.R., Thalmann, D., 1997. A model of human crowd behavior: Group inter-relationship and collision detection analysis, in: Computer animation and simulation, Springer. pp. 39–51.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th international conference on Machine Learning (ICML-10), pp. 807–814.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J., 2017. Plug & play generative networks: Conditional iterative generation of images in latent space, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4467–4477.

Odena, A., 2016. Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583 .

Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier gans, in: Proceedings of the 34th International Conference on Machine Learning - Volume 70, JMLR.org. pp. 2642–2651. URL: http://www.dl.acm.org/citation.cfm?id=3305890.3305954.

Ouyang, Y., Sanchez, V., 2020. Video anomaly detection by estimating likelihood of representations. arXiv preprint arXiv:2012.01468 .

Patino, L., Ferryman, J., 2014. Pets 2014: Dataset and challenge, in: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 355–360. doi:10.1109/AVSS.2014.6918694.

Popoola, O.P., Wang, K., 2012. Video-based abnormal human behavior recognition 2014: A review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42, 865–878. doi:10.1109/TSMCC.2011.2178594.

Pourreza, M., Mohammadi, B., Khaki, M., Bouindour, S., Snoussi, H., Sabokrou, M., 2021a.

G2d: Generate to detect anomaly, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2003–2012.

Pourreza, M., Salehi, M., Sabokrou, M., 2021b. Ano-graph: Learning normal scene contextual graphs to detect video anomalies. arXiv preprint arXiv:2103.10502 .

Prentice, R.L., 1974. A log gamma model and its maximum likelihood estimation. Biometrika 61, 539–544. URL: https://www.doi.org/10.1093/biomet/61.3.539, doi:10.1093/biomet/61.3.539.

Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X., 2018. Pose-normalized image generation for person re-identification, in: Proceedings of the European conference on Computer Vision (ECCV), pp. 650–667.

Qiu, P., Kim, S., Lee, J., Choi, J., 2018. Anomaly detection in a crowd using a cascade of deep learning networks, in: Bhateja, V., Nguyen, B.L., Nguyen, N.G., Satapathy, S.C., Le, D.N. (Eds.), Information Systems Design and Intelligent Applications, Springer Singapore, Singapore. pp. 596–607.

Rabaud, V., Belongie, S., 2006. Counting crowded moving objects, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 705–711. doi:10.1109/CVPR.2006.92.

Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 .

Ramachandra, B., Jones, M., Vatsavai, R.R., 2020. A survey of single-scene video anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Ranjan, V., Le, H., Hoai, M., 2018. Iterative crowd counting, in: The European Conference on Computer Vision (ECCV), pp. 270–285.

Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., Sebe, N., 2016. Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection. CoRR abs/1610.00307, 1689–1698.

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

Ravanbakhsh, M., Sangineto, E., Nabi, M., Sebe, N., 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1896–1904.

Redmon, J., Farhadi, A., 2018. YOLOv3: An incremental improvement. arXiv e-prints , arXiv–1804.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H., 2016. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 .

Ristani, E., Solera, F., Zou, R.S., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking. CoRR abs/1609.01775.

Ristani, E., Tomasi, C., 2018. Features for multi-target multi-camera tracking and re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6036–6046.

Rodriguez, M., Ali, S., Kanade, T., 2009. Tracking in unstructured crowded scenes, in: 2009 IEEE 12th International Conference on Computer Vision, pp. 1389–1396. doi:10.1109/ICCV.2009.5459301.

Rodriguez, M., Laptev, I., Sivic, J., Audibert, J.Y., 2011a. Density-aware person detection and tracking in crowds, in: 2011 International Conference on Computer Vision, IEEE. pp. 2423–2430.

Rodriguez, M., Sivic, J., Laptev, I., Audibert, J.Y., 2011b. Data-driven crowd analysis in videos, in: 2011 International Conference on Computer Vision, pp. 1235–1242. doi:10.1109/ICCV.2011.6126374.

Rosten, E., Drummond, T., 2006. Machine learning for high-speed corner detection, in: European conference on Computer Vision, Springer. pp. 430–443.

Ryan, D., Denman, S., Fookes, C., Sridharan, S., 2009. Crowd counting using multiple local features, in: 2009 Digital Image Computing: Techniques and Applications, pp. 81–88. doi:10.1109/DICTA.2009.22.

Sabokrou, M., Fathy, M., Hoseini, M., Klette, R., 2015. Real-time anomaly detection and localization in crowded scenes, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) workshops, pp. 56–62.

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., Klette, R., 2018. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding 172, 88–97. URL: http://www.sciencedirect.com/science/article/pii/S1077314218300249, doi:https://www.doi.org/10.1016/j.cviu.2018.02.006.

Saleh, S.A.M., Suandi, S.A., Ibrahim, H., 2015. Recent survey on crowd density estimation and counting for visual surveillance. Engineering Applications of Artificial Intelligence 41, 103–114. URL: http://www.sciencedirect.com/science/article/

pii/S0952197615000081, doi:https://www.doi.org/10.1016/j.engappai.2015.01.007.

Salim, S., Khalifa, O.O., Rahman, F.A., Lajis, A., 2019. Crowd detection and tracking in surveillance video sequences, in: 2019 IEEE International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), IEEE. pp. 1–6.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans, in: Advances in Neural Information Processing Systems, pp. 2234–2242.

Schapire, R.E., Singer, Y., 1999. Improved boosting algorithms using confidence-rated predictions. Machine Learning 37, 297–336. URL: https://www.doi.org/10.1023/A:1007614523901, doi:10.1023/A:1007614523901.

Shao, J., C. Loy, C., Wang, X., 2014. Scene-independent group profiling in crowd, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2219–2226.

Shao, J., Kang, K., Change Loy, C., Wang, X., 2015. Deeply learned attributes for crowded scene understanding, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 4657–4666.

Shao, J., Loy, C.C., Wang, X., 2017. Learning scene-independent group descriptors for crowd understanding. IEEE Transactions on Circuits and Systems for Video Technology 27, 1290–1303. doi:10.1109/TCSVT.2016.2539878.

Shehzed, A., Jalal, A., Kim, K., 2019. Multi-person tracking in smart surveillance system for crowd counting and normal/abnormal events detection, in: 2019 International Conference on Applied and Engineering Mathematics (ICAEM), IEEE. pp. 163–168.

Shen, Z., Xu, Y., Ni, B., Wang, M., Hu, J., Yang, X., 2018. Crowd counting via adversarial cross-scale consistency pursuit, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5245–5254.

Shi, J., Tomasi, C., 1994. Good features to track, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 593–600. doi:10.1109/CVPR.1994.323794.

Sidla, O., Lypetskyy, Y., Brandle, N., Seer, S., 2006. Pedestrian detection and tracking for counting applications in crowded situations, in: 2006 IEEE International Conference on Video and Signal Based Surveillance, p. 70. doi:10.1109/AVSS.2006.91.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, pp. 568–576.

Sindagi, V.A., Patel, V.M., 2018. A survey of recent advances in cnn-based single image crowd counting and density estimation. Pattern Recognition Letters 107, 3–16.

Singh, K., Rajora, S., Vishwakarma, D.K., Tripathi, G., Kumar, S., Walia, G.S., 2020. Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. Neurocomputing 371, 188–198.

Sjarif, N., Shamsuddin, S., Hashim, S., 2012. Detection of abnormal behaviors in crowd scene: a review. International Journal of Advances in Soft Computing and its Applications 4, 1–33.

Sjarif, N. N. A.and Shamsuddin, S.M.H., Hashim, S.Z.M., Yuhaniz, S.S., 2011. Crowd analysis and its applications, in: ICSECS (1), pp. 687–697.

Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statistics and Computing 14, 199–222.

Soomro, K., Zamir, A.R., Shah, M., 2012. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 .

Sun, D., Yang, X., Liu, M., Kautz, J., 2017. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. CoRR abs/1709.02371, 8934–8943. URL: http://www.arxiv.org/abs/1709.02371.

Swathi, H., Shivakumar, G., Mohana, H., 2017. Crowd behavior analysis: a survey, in: 2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT), IEEE. pp. 169–178.

Tang, N.C., Lin, Y.Y., Weng, M.F., Liao, H.Y.M., 2015. Cross-camera knowledge transfer for multiview people counting. IEEE Transactions on Image Processing 24, 80–93. doi:10.1109/TIP.2014.2363445.

Tavares, H.L., Neto, J.B.C., Papa, J.P., Colombo, D., Marana, A.N., 2019. Tracking and re-identification of people using soft-biometrics, in: 2019 XV Workshop de Visão Computacional (WVC), IEEE. pp. 78–83.

Tomasi, C., Detection, T.K., 1991. Tracking of point features. International Journal of Computer Vision , 137–154.

Tripathi, G., Singh, K., Vishwakarma, D.K., 2018. Convolutional neural networks for crowd behaviour analysis: a survey. The Visual Computer , 1–24.

Tzutalin, D., 2017. Labelimg. Available at https://github.com/tzutalin/labelImg. [Accessed 13 July 2019].

University of Minnesota, 2006. Detection of unusual crowd activity. Available at http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi. [Accessed 21 November 2017].

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A., 2008. Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th international conference on Machine learning, pp. 1096–1103.

Viola, P., Jones, M.J., Snow, D., 2003. Detecting pedestrians using patterns of motion and appearance, in: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 734–741 vol.2. doi:10.1109/ICCV.2003.1238422.

Vishwakarma, S., Agrawal, A., 2013. A survey on activity recognition and behavior understanding in video surveillance. The Visual Computer 29, 983–1009. URL: https://www.doi.org/10.1007/s00371-012-0752-6, doi:10.1007/s00371-012-0752-6.

Vu, H., Nguyen, T., Le, T., Luo, W., Phung, D., 2019. Robust anomaly detection in videos using multilevel representations, in: Van Hentenryck, P., Zhou, Z. (Eds.), Proceedings of AAAI19-Thirty-Third AAAI conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence (AAAI). pp. 5216–5223. doi:10.1609/aaai.v33i01.33015216.

Wang, X., Ma, X., Grimson, E., 2007. Unsupervised activity perception by hierarchical bayesian models, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. doi:10.1109/CVPR.2007.383072.

Wang, X., Shrivastava, A., Gupta, A., 2017. A-fast-rcnn: Hard positive generation via adversary for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2606–2615.

Wu, S., Yang, H., Zheng, S., Su, H., Fan, Y., Yang, M., 2017. Crowd behavior analysis via curl and divergence of motion trajectories. International Journal of Computer Vision 123, 499–519. URL: https://www.doi.org/10.1007/s11263-017-1005-y, doi:10.1007/s11263-017-1005-y.

Wu, Y., Lim, J., Yang, M.H., 2013. Online object tracking: A benchmark, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418.

Xu, D., Yan, Y., Ricci, E., Sebe, N., 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding 156, 117–127. Image and Video Understanding in Big Data.

Xu, M., Yu, X., Chen, D., Wu, C., Jiang, Y., 2019. An efficient anomaly detection system for crowded scenes using variational autoencoders. Applied Sciences 9, 3337.

Yang, H., Cao, Y., Wu, S., Lin, W., Zheng, S., Yu, Z., 2012. Abnormal crowd behavior detection based on local pressure model, in: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–4.

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C., 2020. Deep learning for person re-identification: A survey and outlook. arXiv preprint arXiv:2001.04193 .

Yogameena, B., Nagananthini, C., 2017. Computer vision based crowd disaster avoidance system: A survey. International Journal of Disaster Risk Reduction 22, 95–29. URL: http://www.sciencedirect.com/science/article/pii/S2212420916302916, doi:https://www.doi.org/10.1016/j.ijdrr.2017.02.021.

YouTube, 2020. Youtube terms of service: Protecting your identity. Available at https://support.google.com/youtube/answer/2801895?hl=en&ref_topic=9386940. [Accessed 13 July 2020].

Yu, J., Lee, Y., Yow, K.C., Jeon, M., Pedrycz, W., 2021. Abnormal event detection and localization via adversarial event prediction. IEEE Transactions on Neural Networks and Learning Systems , 1–15doi:10.1109/TNNLS.2021.3053563.

Zhan, B., Monekosso, D.N., Remagnino, P., Velastin, S.A., Xu, L., 2008. Crowd analysis: a survey. Machine Vision and Applications 19, 345–357. URL: https://www.doi.org/10.1007/s00138-008-0132-4, doi:10.1007/s00138-008-0132-4.

Zhang, C., Kang, K., Li, H., Wang, X., Xie, R., Yang, X., 2016a. Data-driven crowd understanding: A baseline for a large-scale crowd dataset. IEEE Transactions on Multimedia 18, 1048–1061. doi:10.1109/TMM.2016.2542585.

Zhang, F., Xue, W., Cui, L., Zhu, G., 2018a. A crowd anomaly behavior detection algorithm, in: 2018 International Conference on Audio, Language and Image Processing (ICALIP), pp. 457–463. doi:10.1109/ICALIP.2018.8455412.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N., 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 5907–5915.

Zhang, X., Yu, Q., Yu, H., 2018b. Physics inspired methods for crowd video surveillance and analysis: a survey. IEEE Access 6, 66816–66830.

Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y., 2016b. Single-image crowd counting via multi-column convolutional neural network, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 589–597).

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X., 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1077–1085.

Zhao, M., Zhong, J., Cai, W., 2018. A role-dependent data-driven approach for high-density crowd behavior modeling. ACM Transactions on Modeling and Computer Simulation (TOMACS) 28, 1–25.

Zhao, T., Nevatia, R., 2003. Bayesian human segmentation in crowded situations, in: Computer Vision and Pattern Recognition (CVPR), 2003. Proceedings. 2003 IEEE Computer Society Conference on, IEEE. pp. II–459.

Zhao, T., Nevatia, R., Wu, B., 2008. Segmentation and tracking of multiple humans in crowded environments. IEEE Transactions on Pattern Analysis and Machine Intelligence 30, 1198–1211. doi:10.1109/TPAMI.2007.70770.

Zhou, B., Wang, X., Tang, X., 2012. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2871–2878. doi:10.1109/CVPR.2012.6248013.

Zhu, W., Xiang, X., Tran, T., Xie, X., 2017. Adversarial deep structural networks for mammographic mass segmentation. arXiv preprint arxiv:1612.05970 .

Zitouni, M.S., Bhaskar, H., Dias, J., Al-Mualla, M., 2016. Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques. Neurocomputing 186, 139–159. URL: http://www.sciencedirect.com/science/article/pii/S092523121502041X, doi:https://www.doi.org/10.1016/j.neucom.2015.12.070.

A Previous Reviews

The most notable survey/review papers in crowd analysis are investigated (Table 15). The selected surveys are amongst the papers published within the last ten years. One of the most cited surveys is written by Zhan et al. (2008); this is one of the first surveys about crowd analysis. This survey mainly reviews crowd analysis techniques based on computer vision approaches, but, other crowd analysis perspectives like sociology, psychology, and computer graphics are also explored. The main topics focused on are crowd density estimation, recognition, tracking, crowd modelling, and event interpretation.

Ko (2008) introduced a survey where the focus was on hardware and software combinations that can help solve surveillance challenges. The advances and strategies used in video surveillance are reviewed in addition to motion analysis, behaviour analysis, biometrics, anomaly detection, and behaviour understanding. Computer vision techniques for crowd analysis are covered by Junior et al. (2010); more specifically, issues such as tracking, crowd density measurement, event inference, validation, and simulation. The main difficulties explored in the survey were density estimation/crowd counting, tracking within a crowd, and higher-level analysis for the understanding of crowd behaviour.

A detailed survey, presented by Candamo et al. (2010), on human behaviour recognition approaches for transportation surveillance had four main focuses. The focuses were categorised into recognition of a single person, multiple persons, person and vehicle interaction, and person and location interaction. Some of the interactions recognised were loitering, fights or attacks, vehicle damage, and deserting personal belongings. State-of-the-art advancements in motion detection, moving objects classification, and tracking is also presented. The review paper Sjarif et al. (2012), investigated state-of-the-art techniques in analysing crowd behaviour between the years of 2000 to 2010. Crowd density measurement, crowd motion recognition, tracking, and crowd behaviour detection are explored in relation to abnormal event recognition. Approaches in pre-processing, object tracking, and event/behaviours detection is evaluated in detail.

Similarly, Popoola and Wang (2012) proposed a review paper that presents the most recent developments in abnormal human behaviour detection from video footage. Past reviews are explored and mentioned while maintaining a focus on the recognition of abnormal behaviour, particularly in video surveillance. Detailed highlights of current methods were presented in a manner such that the main challenges in behaviour analysis are brought to notice.

An attempt is made by Chaquet et al. (2013) to cover the absence of information on the most significant and public video-based datasets for human action and movement recognition. The survey is of great help to researchers who required selection of the most appropriate benchmark datasets for their algorithms. An assessment of the current datasets is provided with the emphasis on ground truth data, scene varieties, actions/human count, and references to published papers utilising these datasets.

Another notable survey is Vishwakarma and Agrawal (2013) where the main focus is the detection in video surveillance. The authors detail pre-processing techniques, object tracking approaches, and activity recognition methods. Recent research relating to activity detection, benchmark datasets, and applications was also been documented. As well as the various methods for action recognition of a single human or a crowd as a whole.

Cristani et al. (2013) aim to review the more noteworthy human behaviour analysis work that combines both video surveillance and Social Signal Processing (SSP) . An investigation on where surveillance and social signalling intersect is documented, as well as how social signalling may aid in the progression of the analysis of human behaviour. Similar to the survey undertaken by Zhan et al. (2008) a more updated survey by Li et al. (2015) explores state-of-the-art techniques in crowd motion pattern learning, crowd behaviour and activity analysis, and crowd anomaly recognition. The paper explores many aspects of crowd analysis such as current models, widespread algorithms, protocols for evaluation, and system performance. Also documented, are the available evaluation datasets, research problems, and promising future work.

A literature review compiled by Afsar et al. (2015) investigates 193 papers from the years of 2000 to 2014 about visual detection of human behaviour. The review categorised these papers into three topics: techniques for detection, datasets, and applications. The review further sub-categorised each topic into a deeper classification where detection techniques were divided into initialisation, tracking, pose estimation, and recognition. Applications such as human detection, abnormal behaviour detection, activity recognition, modelling, and pedestrian detection were investigated. Additionally, eight datasets were listed that can assist future researchers in their human behaviour detection systems.

Zitouni et al. (2016) study a more specific topic; the past seven years of research on crowd modelling techniques are explored. The target of the paper is to make recommendations based on the general features of the techniques instead of explicit algorithms. The survey also presents a comparison of current methods using public crowd datasets based on quantitative and qualitative features. Kok et al. (2016) take in a non-typical approach where an investigation is applied to crowd behaviour analysis based on a physics and biology perspective. The authors examine these two sciences taking into consideration previously ignored areas, as well as the explored areas for analysing crowd behaviour. Additionally, the authors discuss the essentiality of merging both biology and physics sciences in computer vision.

In 2017 four prominent surveys have been published each of which has an explicit focus. Firstly, Convolutional Neural Network (CNN) approaches for crowd counting and density estimation are surveyed by Sindagi and Patel (2018). Furthermore, an evaluation and comparison between these CNN approaches and earlier hand-crafted methods are noted, as well as newly published datasets. Secondly, Yogameena and Nagananthini (2017) have aimed their survey at the investigation of current developments and approaches in crowd disaster analysis that can create a stable Computer Vision-Crowd Disaster Avoidance System (CV-CDAS). One of the significant influences is behaviour analysis, which is explored in detail, also noted is an evaluation of the benchmark datasets. The third published survey has two main focuses: crowd statistics and behaviour understanding (Grant and Flynn, 2017). Crowd counting and density estimation methods are first investigated, then research related to crowd behaviour understanding is presented. Tracking approaches are also surveyed, as well as crowd behaviour video datasets. Lastly, Swathi et al. (2017) present a less extensive review on crowd behaviour analysis, but it is a good guide to new researchers. The review includes basic framework architectures of video surveillance and crowd analysis. Basic terminology used in crowd analysis are also described based on a accumulation of different definitions given by various authors.

Newer reviews such as Haghani and Sarvi (2018), Zhang et al. (2018b), and Tripathi et al. (2018) address more novel approaches for crowd behaviour analysis. Haghani and Sarvi (2018) have evaluated almost 150 studies in relation to minimisation of crowd disaster and evacuation planning. It is a very extensive review including keywords such as "crowd motion", "emergency evacuation", "animals", and "walking behaviour". The review is very diverse regarding collection of data related to crowd analysis such as: animal experimentation, human controlled experimentation, virtual reality experimentation, evacuation experimentation, and natural disaster evaluation. One of the authors' most interesting findings is crowd behaviour analysis studies often have conflicting definitions of basic terminology. Authors have contradicting evidence about their data, and the evaluation metrics applied are biased towards the research. Zhang et al. (2018b) surveys physics inspired methods for crowd analysis and surveillance. Due to the fact that crowds exhibit features like velocity, direction, energy and force all being based on physics. The methods examined by the authors are divided into three classifications: fluid dynamics, interaction force, and complex crowd motion systems. Similar to other review papers, benchmark datasets and unresolved areas are deliberated. Lastly, Tripathi et al. (2018) provide a more distinct review on crowd behaviour analysis methods based on convolutional neural networks (CNN). Reviewed are topics such as the evolution of CCN in the field of crowds behaviour, challenges in this field, CNN methods previously applied by researchers, and applicable datasets. A significant finding of this review is high-density level crowds are still difficult to analyse in regards to detecting and tracking objects, crowd counting, and detecting anomaly.

Table 15: Notable previous surveys in chronological order

Survey Title	**Author and Year**
A survey on behaviour analysis in video surveillance for homeland security applications	(Ko, 2008)
Crowd analysis: a survey	(Zhan et al., 2008)
Crowd analysis using computer vision techniques	(Junior et al., 2010)
Understanding transit scenes: A survey on human behaviour-recognition algorithms	(Candamo et al., 2010)
Detection of abnormal behaviours in crowd scene: a review	(Sjarif et al., 2012)
Video-based abnormal human behaviour recognition-A review	(Popoola and Wang, 2012)
A survey of video datasets for human action and activity recognition	(Chaquet et al., 2013)
A survey on activity recognition and behaviour understanding in video surveillance	(Vishwakarma and Agrawal, 2013)
Human behaviour analysis in video surveillance: A social signal processing perspective	(Cristani et al., 2013)
Crowded scene analysis: A survey	(Li et al., 2015)
Automatic visual detection of human behaviour: a review from 2000 to 2014	(Afsar et al., 2015)
Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques	(Zitouni et al., 2016)
Crowd behaviour analysis: A review where physics meets biology	(Kok et al., 2016)
A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation	(Sindagi and Patel, 2018)
Computer Vision based Crowd Disaster Avoidance System: A Survey	(Yogameena and Nagananthini, 2017)
Crowd Scene Understanding from Video: A Survey	(Grant and Flynn, 2017)
Crowd behavior analysis: a survey	(Swathi et al., 2017)
Crowd behaviour and motion: Empirical methods	(Haghani and Sarvi, 2018)
Physics inspired methods for crowd video surveillance and analysis: a survey	(Zhang et al., 2018b)
Convolutional neural networks for crowd behaviour analysis: a survey	(Tripathi et al., 2018)

B Crowd Counting / Density Estimation

Traditional approaches for crowd counting and density estimation are reviewed below.

B.1 Direct Approach

A direct approach for crowd counting tries to identify every single person within a scene along with their corresponding position. As long as the segmentation process is correctly applied, the number of people is easy to obtain. The segmentation process can differ for each of the methods; some methods segment the whole contour of the body (head, shoulders, arms, and legs), while others efficiently segment the Ω-form of the human (head and shoulders). The problems that arise with the application of this method are occlusion handling and handling high-density crowds. Some of the sub-areas to be considered when using the direct approach include model-based methods and trajectory-based clustering methods (Saleh et al., 2015).

B.1.1 Model-based Approach

Viola et al. (2003) present a pedestrian detection system; their system takes advantage of both image appearance data and motion data by combining them to detect a person who is walking. The image appearance data used is based on feature extraction using an integral image. The authors utilise two consecutive video frames using a detection-based algorithm, which is fast and efficient. A detector is then trained using AdaBoost (Schapire and Singer, 1999). This approach is applicable to many difficult scenarios such as low-resolution images, and crowds in bad weather conditions like rain and snow which cause low visibility. The system is trained and tested on a dataset of scenes from the street created by the authors. The pedestrians were highlighted with a box in each frame. The dataset had eight sequences, with an approximation of 2000 frames for each sequence. Six of the sequences were used for training the detection of both dynamic and static pedestrians, while the other two sequences were used for testing. When the algorithm was tested on the two sequences it achieved very low false-positive rates; the best result being 1 in 400,000 false positives. In addition, a good detection rate was attained by the algorithm with the highest result of 80%.

Research by Lin et al. (2001) developed a system that can use a single image to make an approximation of the number of people in a crowd, even if the background of the scene is of complex nature. The system approaches this by identifying the contour of people's heads. To extract the features of any head-shaped contour, the authors propose using the Haar Wavelet Transform (HWT) function (Chapelle et al., 1999). Additionally, the system will determine the input features as either head or not using a support vector machine (SVM) with three stages:

pre-processing the image, extracting features, and support vector classifying. To precisely estimate the size of the crowd, a technique using perspective transformation (also known as imaging transformation) is utilised. For system testing, the authors developed a crowd simulation model world with 125 human-shaped puppets. All the ground truth data about the model world is noted taking into consideration the crowd size and angle view. In this model world, the system showed an overall accuracy level of 90% - 95% with a reduction in accuracy as the size of the crowd grows.

Research presented by Zhao et al. (2008) present an approach within a Bayesian framework that can model multiple partially occluded humans. This model-based stochastic approach has the advantage of not needing a person to be un-occluded when entering the scene, but only requires the visibility of the head and shoulder region. The overview of the approach is shown in Figure 110. The method basically starts with blob boundaries detection, then the canny edge detection algorithm is applied to extract the edges of the subjects, the head and shoulder model is applied afterwards. Lastly, using edge intensity, humans can be reliably detected. With the use of a sampling method, data-driven Markov chain Monte Carlo (DDMCMC), a configuration that can adequately clarify the foreground mask is estimated. The approach is tested on outdoor and indoor footage each including occlusion events. In the outdoor testing, the dataset comprises of 33 people going in opposite directions with 20 occlusions, 9 of them considered as heavy occlusion. The results of the outdoor testing were 98.13% detection rate and false detection rate of 0.27%. The detection and false-alarm rate for the indoor dataset, Context-Aware Vision using Image-based Active Recognition (CAVIAR, 2003) were not detailed, but the paper signifies the approach can show promise if integrated with an improved background and shadow model. The work proposed by Ge and Collins (2009) is an extension of the approach, presenting an improved technique with the use of shaping models that are more practical and flexible.

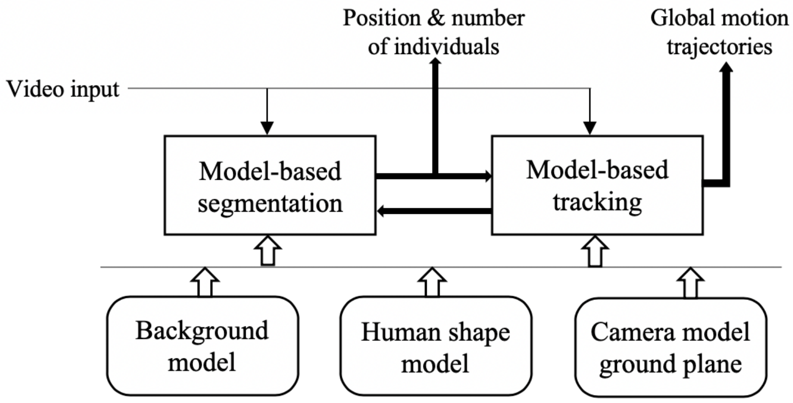


Figure 110: Overview of approach. Adapted from (Zhao et al., 2008)

B.1.2 Trajectory-based clustering Approach

Brostow and Cipolla (2006) investigate an approach to distinguish separate movement in a crowd by using an unsupervised Bayesian clustering algorithm. The simple idea of the approach assumes that a pair of points that are parallel in movement probably belong to the same object. The authors characterise a moving object by extracting low-level features of an image, which are clustered using probabilistic behaviour. An advantage of this approach is that there is no requirement for training to achieve its goal. The paper considers tracing two features: both Rosten and Drummond (2006) features and Tomasi and Detection (1991) features. Additionally, to track the features in two frames hierarchical optical flow (further explained in Section 3.4.1) is applied. To test the approach, 10 sequences were used each of which is between three seconds to one hour long. A comparison between the authors' results and the previously discussed approach by Zhao et al. (2008) shows that the detection rate is lower with a result of 94%, and the false detection rate was significantly higher rising up to 22.9%. Moreover, failure indication and false detection were noticed if the system is presented with vigorous arm movement.

Rabaud and Belongie (2006) presented an approach to segment an individual within a crowded scene by the use of the individual's motion within multiple occurrences. Similarly to Sidla et al. (2006), the authors approached this by using a Kanade-Lucas-Tomasi (KLT) (Tomasi and Detection, 1991) tracker with a more parallelised manner. The KLT tracker is an algorithm used to extract features for multiple purposes such as camera motion estimation, video stabilisation, or object tracking. Using this algorithm for object tracking works best with objects that do not change shape or formation. The tracker uses spatial intensity data to aim the search in the direction of finding the best match. With the use of this tracker, a large set of low-level features were extracted in an enhanced mean. Additionally, with the use of spatial and temporal conditioning, the trajectories are filtered to recognise the number of moving objects within a scene. The authors tested their approach on three datasets of real-world imagery: a USC dataset (Zhao and Nevatia, 2003), the author's dataset: LIBRARY, and CELLS dataset with footage of red blood cells movement. The approach shows reasonable occlusion handling, but when demonstrated with shared motion between interacting objects it caused the trajectories to be merged inaccurately. Detailed results of the testing on the datasets are noted in the paper, with an average error rate of 10%, 6.3% and 22% for the USC, LIBRARY, and CELLS datasets respectively.

B.2 Indirect Approach

Compared to the direct method, the more efficient approach is the indirect one. The approach does not try and detect a person directly, but to represent a crowd, it typically extracts multiple local and holistic features from foreground images. The extraction of these features has proven to be more proficient than person detection. There are multiple variants to the indirect approach including pixel-based, texture-based, and corner point-based approaches (Saleh et al., 2015).

A pedestrian detection and tracking system featured by Sidla et al. (2006) calculate points of interest and applies motion prediction. Furthermore, the system finds specific shape information for human detection and implements texture feature extraction for human recognition. The shape chosen to represent humans is a Ω-like shape. The authors detect this shape with a masking filter over the region of interest. A description of each person is deduced from the use of a co-occurrence matrix feature vector by using the Kalman filter (Kalman, 1960). Additionally, KLT tracking points (Tomasi and Detection, 1991) are used. In comparison to more traditional algorithms, the KLT tracker examines far less probable matches between images. While the Kalman filter is used to make an educated guess about the next step the system will make. The linear-quadratic estimation (LQE) algorithm observes a collection of measurements that are presented with noise and other inaccuracies. Over time, the algorithm finds current estimates of variables by approximating the joint probability distribution over the unknown variables for every time-frame. The pedestrian detection approach is tested on two scenarios indoor and outdoor; the indoor scenario makes use of video footage from an underground platform. The results start off with an absolute mean error of 10% over a built-up time of 240 seconds, but as the time interval increases the results decrease reaching a result of 2% over one hour. Although not detailed, the authors claim that similar results were shown when tested in the outdoor scenario.

B.2.1 Pixel-based Approach

A neural-based system that can identify overcrowding in a specific setting is investigated by Cho and Chow (1999); Cho et al. (1999). More specifically the authors targeted platforms in underground stations, and the targeted platforms of their research were the Mass Transit Railway (MTR) stations located in Hong Kong. The system used CCTV imagery which is passed through pre-processing techniques that map the visual data to a two-dimensional feature space by using the extracted features. A visual representation of an overview of the system is shown in Figure 111. The feature extraction targets low-level features with an assumption that a correlation exists between the crowd level and the segmented regions where there is a considerable amount of movement. For each image given, three features are extracted: length of crowd edges, the density of crowd objects and the density of the background. The neural network takes the extracted feature coefficients as its input. A hybridising of the Least Squares (LS) algorithm and a global optimisation method is used as the system's learning algorithm to classify the crowd. The authors test their system using two-hybrid algorithms that combine the LS algorithm with both random search algorithm and Simulated Annealing (SA) algorithm.

The best Sum Squared Error (SSE) for learning is the result of using the LS and SA algorithm with a value of 1.189 and an estimation accuracy result of 94.36%, but with a substantially longer CPU running time than the LS and random search algorithm (Cho and Chow, 1999; Cho et al., 1999).

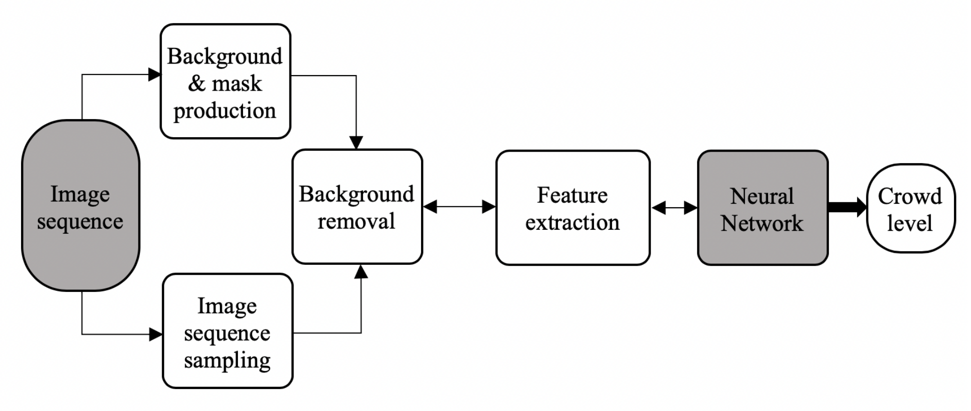


Figure 111: Overview of neural based crowd monitoring system. Adapted from (Cho et al., 1999)

Additionally, Cho et al. (1999) enhance their system with another neural-based crowd estimation hybrid algorithm. The method is a cross-over between the LS algorithm and Genetic Algorithm (GA). The same neural network topology is used, which consists of three inputs, 15 hidden neurons, and one output neuron. So far, the fastest algorithm remains the hybrid of LS and random SA. The given hybrid algorithm has very close results to the initial algorithm using LS and SA with an estimation accuracy result of 93.8% and 94.36% respectively, but with a CPU running time decreased to less than half the running time of LS/SA.

Tang et al. (2015) presented a system to count the number of people in a crowd using a two-pass regression framework; several cameras are used to bring diverse views of the same crowd. With the use of different views, the system can gather corresponding data to enhance the performance and crowd counting process. The authors first tackle the problems of estimating the crowd count and normalising the visual feature perspective, considering them to be one learning problem. Subsequently, the paper presents an algorithm that receives multiple views of a crowd and matches the groups from each view. Lastly, the authors detail the regressors used in the system: where one of the regressors uses the extracted features from the intra-camera images given to count the crowd, and the other regressor determines the remaining count with respect to the inter-camera predictions conflict. The results were presented using mean absolute error (MAE) with an overall result of 3.26 using the first-pass regression approach (FPR), and 2.52 using the two-pass regression framework (TPR). The data was trained and tested on four levels of crowd density using the PETS 2009 benchmark dataset (Ferryman and Shahrokni, 2009). The data was annotated as: sparse (few people, minimum occlusions), medium (more than a few people, medium occlusions), heavy (dense crowd, full occlusions), and mixed (a combination of the three). The paper also noted comparisons with other baseline approaches in which the authors' suggested approach outperforms them.

B.2.2 Texture-based Approach

Chan et al. (2008) present an approach for unstructured crowd estimation without the use of tracking approaches or explicit object segmentation. Their contributions are threefold: crowd counting while preserving the privacy of people in the crowd, validating the approach by using a dataset comprising of 49,885 pedestrian instances, and finally, the robustness of the approach is shown by testing on an hour-long video. Initially, the system used a combination of dynamic textures to segment the crowd into multiple motion features. The system used a Gaussian process for counting the number of people. It was trained on 800 frames and tested on 1200 frames. The best results presented for both directions (away and towards) were MSE (Mean-Squared-Error) values of 4.181 (away) and 1.29 (towards). This result was achieved when using all the extracted features (segmentation, internal edge, and texture features).

Subsequently, Chan and Vasconcelos (2009) took another route and investigated a standard Poisson regression model in a Bayesian setting. The authors initially developed a closed-form approximation to the predictive distribution of the Bayesian Poisson regression (BPR) model. The predictive distribution was then kernelised and through kernel functions the representation of non-linear log-mean functions was admissible. To enhance the hyper-parameter of the kernel function an estimated marginal likelihood function was developed and used to show its relation to a Gaussian process with a special non-i.i.d. (non-independent and identically distributed) noise term. The authors experimented using the crowd video database (Chan et al., 2008), 1200 frames were used for training and 2600 frames for testing. The best results were obtained where the trends in the log-mean function were modelled using a kernel consisting of two radial basis functions (RBF) (Prentice, 1974). The results documented were MSE of 2.4675 ("Away") and 2.0246 ("Towards").

A crowd counting approach that uses local features instead of holistic features is proposed in (Ryan et al., 2009). The authors use a foreground subtraction technique, and the local features are extracted with respect to blob segments. The number of people in each blob segment is estimated so that the accumulation of all the segments in the scene is the scene estimation. The authors tested their approach on two classifiers: a neural network classifier and a linear model classifier. The training set, which consisted of 160 frames, used for the classifier was manually annotated with ground truth data (number of segmented blobs). Additionally, the neural network was trained consecutively five times and the median MSE was noted. The lowest MSE result occurred using the linear model classifier with a result of 3.065, this was tested on 1200 frames. The authors claim the neural network would have presented better results if the training data were larger.

B.2.3 Corner point-based Approach

Kong et al. (2006) use a feed-forward neural network in a viewpoint invariant learning-based method to map the connection between the number of pedestrians and the feature histogram extracted from low-level features. Instead of simple features, the authors use feature histograms because they consider it to be more accurate in terms of pedestrian counting and better with noise handling. The system extracts background and edge features, fuses them together, and then normalises them with respect to perspective projection and camera orientation. These features are then used to train a supervised feed-forward neural network. The authors test their suggested method using footage from multiple venues that include different camera positioning. They present their result in a graphical format to display the performance and prospect of the algorithm but error evaluation measurements were not provided.

A multi-output regression model is presented in Chen et al. (2012) where the model was automated to learn the functional mapping between multi-dimensional structured output and interdependent low-level features. Even with diverse environments, the model was capable of counting people by finding the intrinsic importance of multiple features. The outline of the model is described in four steps: initially, a perspective normalisation map is deduced by using the Chan et al. (2008) technique. Subsequently, for each cell region, low-level features such as foreground, edge, and texture are extracted from the training set. Next, the extracted features of each cell are used to create intermediate feature vectors, which are connected together to form a single feature vector. Lastly, the resulting single feature vector and the intermediate feature vectors are paired for the training of a multi-output regression model based on multiple variants of ridge regression. The authors tested their approach on two datasets: the UCSD dataset (Chan et al., 2008) and their own Mall dataset. The authors followed the standard Train/Test partitioning of the UCSD dataset but for the Mall dataset the authors chose 800 frames for training and 1200 frames for testing; the experiment showed an MSE result of 8.08 and 15.7 for each dataset respectively.

An enhanced crowd counting method proposed by Liang et al. (2014) mainly uses feature points. The goal of the approach was to extract crowd characteristics such as orientation and count. The authors present a three-frame difference algorithm, shown in Figure 112, to find the foreground of only entities that present movement. The extracted foreground was then used for the detection of feature points. The method makes use of the SURF (Speeded Up Robust Feature) algorithm with additional adjustments applied to make the algorithm more robust. Furthermore, after the removal of non-motion feature points, an enhanced clustering algorithm DBSCAN (Density-Based Spatial clustering of Application with Noise) (Ester et al.,

1996) was used to cluster the remaining feature points. The feature points are tracked using the combination of local optical flow (Lucas and Kanade, 1981) (further explained in Section 3.4.1) and Hessian matrix algorithm to determine the crowd flow orientation. Additionally, a support vector regression machine is trained with extracted eigenvectors for crowd counting. The testing of this algorithm displays improved results when compared to other approaches. The authors noted evaluation metrics when testing on the PETS dataset with respect to both crowd flow orientation and crowd counting results. As for crowd counting, four different video sequences with different densities (low, medium, high, combination) are documented with a mean absolute error (MAE) of 1.01%, 1.17%, 4.33%, and 1.39% respectively.

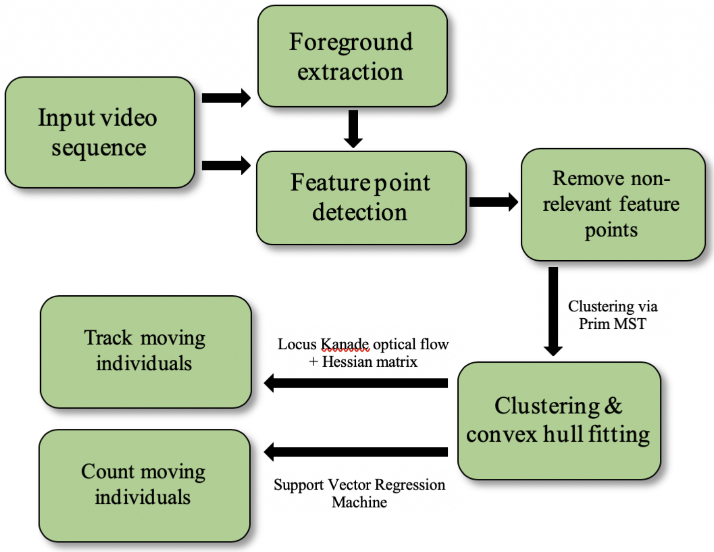

Foreground extraction
Input video sequence
Feature point detection
Remove non-relevant feature points
Clustering via Prim MST
Track moving individuals
Locus Kanade optical flow + Hessian matrix
Clustering & convex hull fitting
Count moving individuals
Support Vector Regression Machine

Figure 112: The three-frame difference algorithm. Adapted from (Liang et al., 2014)

C Tracking / Person Re-Identification

C.1 Contextual Approach

An early paper written by Javed et al. (2008) argues that pedestrians often use similar pathways when walking; the authors use this phenomenon to form a connection between the travelled paths. The suggested algorithm learns the space-time cues and therefore learns the inter-camera connection; these are used to constrain a relationship between cameras. Moreover, with the use of kernel density estimation, the relationships are modelled as probability density functions of space-time variables such as entrance/exit locations, velocity, and transition times. Javed et al. (2008) suggest that objects moving from one camera into another often present appearance alterations. This can be managed with the use of a brightness transfer function between camera pairs that lie in a low dimensional subspace. The probabilistic principal component analysis is used to train the algorithm to learn this subspace. With the use of cues such as location and appearance, a maximum likelihood (ML) estimation framework is implemented for tracking. The algorithm is tested on real-world footage to validate near real-time implementation of the proposed algorithm. However, quantitative results are not documented.

For re-identification using multiple cameras, Gandhi and Trivedi (2007) use a Panoramic Appearance Map (PAM). The approach extracts features from all the footage, taken from multiple cameras, which can view the targeted object. These features are combined to create a single signature. To find the position of the intended object multiple-camera triangulation is used to place a cylinder-shaped model around the location of the object. The panoramic map is created with the horizontal axis representing the azimuth angle, taking into consideration real-world coordinates. Meanwhile, the vertical axis denotes the height of the object with respect to the ground plane. With the use of extracted colour information from different maps, a comparison can be made to determine probable object matches. Gandhi and Trivedi (2007) proposed to do this comparison with the use of weights to the sum of squared differences. For the approach to work properly, there is a requirement that three or more cameras simultaneously view the object. Furthermore, 3D positioning and calibration of the cameras is a must to ensure re-identification.

3D information extracted from footage by various cameras is used for a surveillance system developed by Baltieri et al. (2011) that can detect, track, and re-identify individuals. The approach is built on three key modules: detection of an object, short-term tracking, and long-term tracking. The detection module merges information extracted from all camera views to detect an object and find its location on the ground plane. 3D Marked Point Process model takes two pixel-level features as its input and can then approximate the location and height of the object with the use of a stochastic optimisation framework. For short-term tracking, the authors use a Kalman filter to track individuals taking into use the detection results achieved. Local matching is achieved with the use of geometrical and spatial data. Lastly, long-term tracking finds the trajectories corresponding to the same object and then matches and combines them together for re-identification. The authors evaluate their detection results using two datasets: PETS (Ferryman and Shahrokni, 2009) and EPFL Terrace indoor (Berclaz et al., 2011), with a total error rate (TER) of 10% and 7% respectively. The TER is the summation of the rates of missed detection, false detection, and multiple instances. The long-term tracking was evaluated using precision and recall values with results of 72.73% and 88.8% respectively.

C.2 Non-Contextual Approach

A more novel method, by Bazzani et al. (2010), presented an identification signature named Histogram Plus Epitome (HPE). The method extracts features from multiple images of a human then the features are concentrated to develop the signature. It begins by processing multiple images, preferably from a single-camera, to obtain silhouettes of the body. Images that are considered redundant or outliers are removed using unsupervised Gaussian clustering technique (Figueiredo and Jain, 2002). The human appearance is then described using two complementary features: global and local. The global appearance features are represented using an HSV (hue, saturation and value) histogram, while the local features are encoded through epitomic analysis, which uses recurring local patches. Finally, appearance matching is implemented through a weighted sum of feature similarities. Although quantitative results are not noted, experiments are applied to two datasets: i-LIDS (Advanced Video and Signal based Surveillance, 2007) and ETHZ (Ess et al., 2007). The authors claim their method has better results in comparison to the best performing technique at the time. Additionally, occlusions and crowded scenes are handled well with this method.

Bazzani et al. (2013) and Farenzena et al. (2010) applied Symmetry-Driven Accumulation of Local Features (SDALFs) to distinguish the appearance of an object with the use of visual cues. The descriptor is constructed with symmetry-driven appearance-based features combined with a simple distance minimisation technique for object matching. The approach begins by localising meaningful body parts such as head, upper body, and lower body leading to the removal of unnecessary background data. The localised parts are used to extract three corresponding appearance characteristics. The first makes use of a weighted HSV histogram to encode global chromatic substance. The second uses Maximally Stable Colour Regions (MSCR) to encode the colour displacement per-region. Finally, a per-patch similarity analysis technique is employed to approximate the Recurrent Highly Structured Patches (RHSP). The authors apply the SDALFs method for both re-identification and multiple target tracking. Experimental results are shown with the use of the benchmark dataset (CAVIAR, 2003) for testing, more specifically the shopping centre scene. The authors show comparison against other tracking descriptors with detailed evaluation metrics such as false positives, false negatives, and average tracking accuracy with the results of 0.0608, 0.1852, and 0.4567 respectively (Bazzani et al., 2013). The results note significant enhancements; additionally, the descriptor is able to handle pose, viewpoint, and illumination changes.

D GANs Experimentation

The following image is a collection of the results produced by the various types of GANs implemented on the MNIST dataset. The method type, the results of the first epoch and the results of the last epoch are displayed in Figures 113,114 and 115. The methods experimented below are the works of: (Makhzani et al., 2015)[1], (Odena et al., 2017)[2], (Hjelm et al., 2017)[3], (Donahue et al., 2016)[4], (Denton et al., 2016)[5], (Mirza and Osindero, 2014)[6], (Goodfellow et al., 2014)[7], (Radford et al., 2015)[8], (Chen et al., 2016)[9], (Bousmalis et al., 2017)[10], (Odena, 2016)[11], (Martin A. and Bottou, 2017)[12], (Gulrajani et al., 2017)[13].

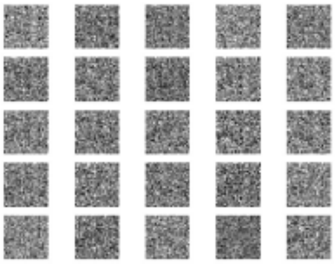
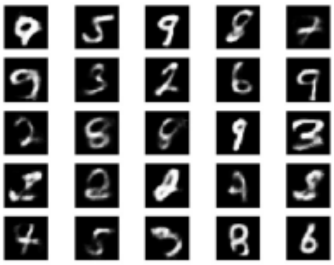
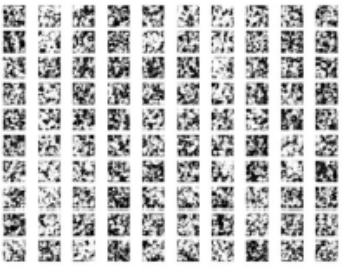
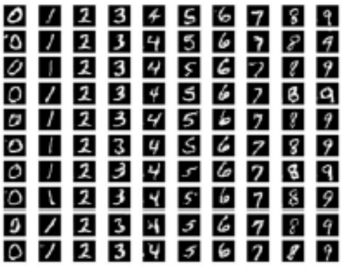
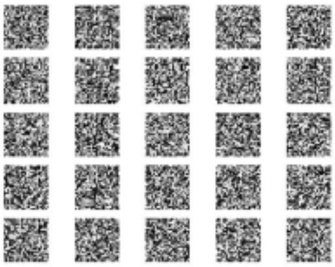
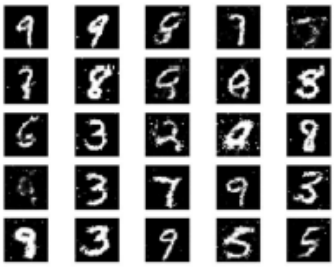
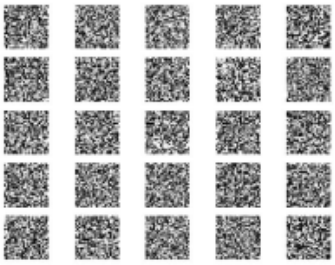
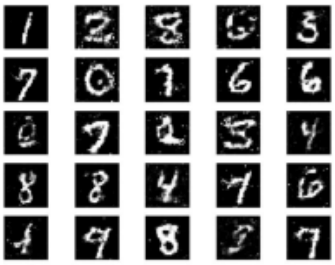
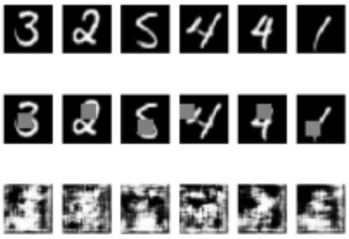
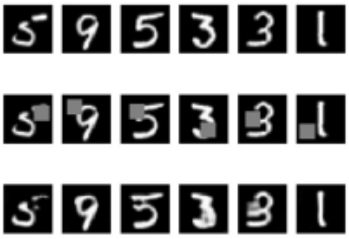
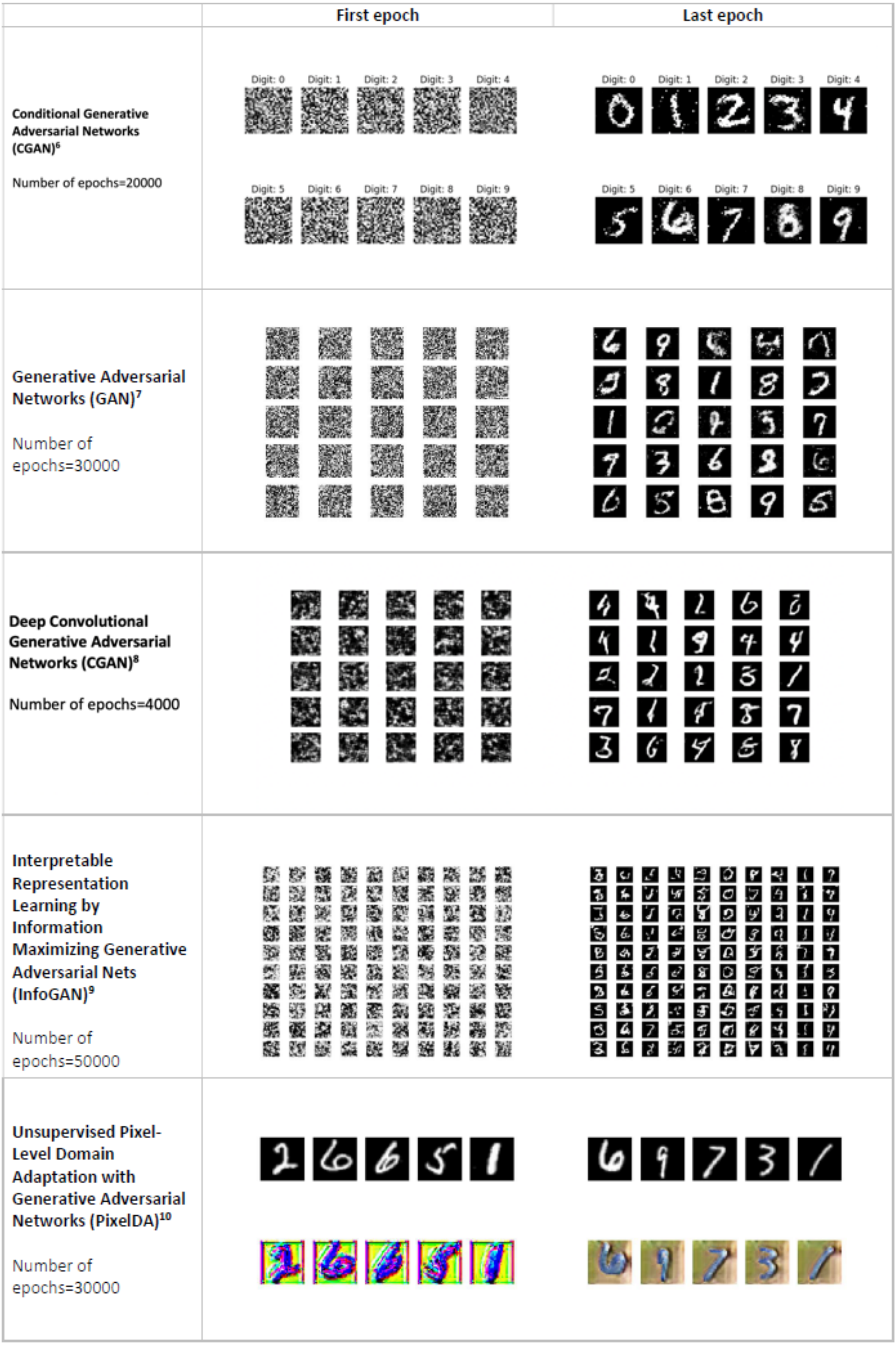
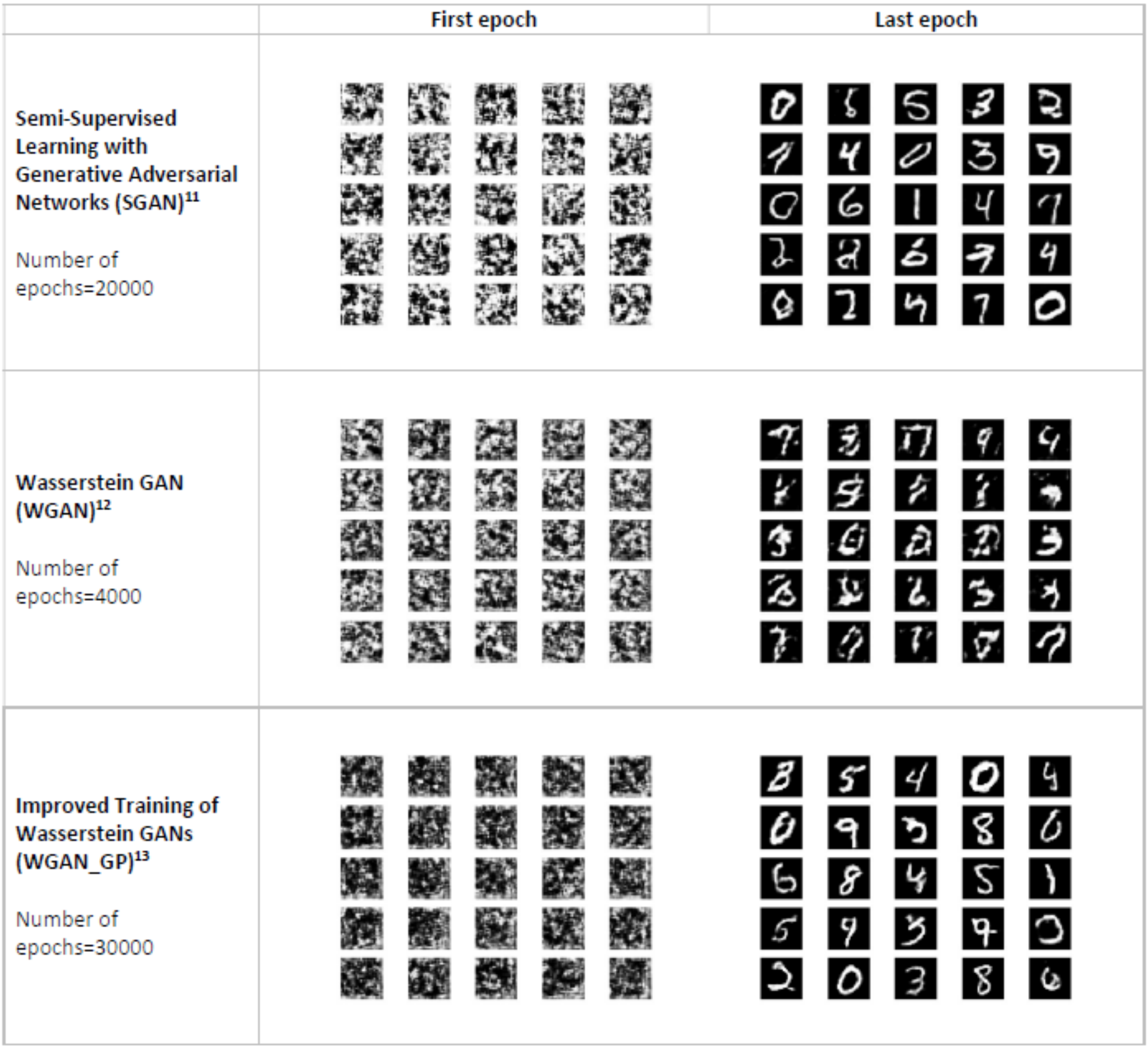
	First epoch	Last epoch
Adversarial Autoencoders (AAE)[1] Number of epochs=20000		
Auxiliary Classifier Generative Adversarial Network (AC-GAN)[2] Number of epochs=14000		
Boundary-Seeking Generative Adversarial Networks (BGAN)[3] Number of epochs=30000		
Bidirectional Generative Adversarial Networks (BiGANs)[4] Number of epochs=40000		
Semi-Supervised Learning with Context-Conditional Generative Adversarial Networks (CC-GAN)[5] Number of epochs=20000		

Figure 113: Results produced from various type of GANs



	First epoch	Last epoch
Conditional Generative Adversarial Networks (CGAN)[6] Number of epochs=20000		
Generative Adversarial Networks (GAN)[7] Number of epochs=30000		
Deep Convolutional Generative Adversarial Networks (CGAN)[8] Number of epochs=4000		
Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets (InfoGAN)[9] Number of epochs=50000		
Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks (PixelDA)[10] Number of epochs=30000		

Figure 114: Results produced from various type of GANs



	First epoch	Last epoch
Semi-Supervised Learning with Generative Adversarial Networks (SGAN)[11] Number of epochs=20000		
Wasserstein GAN (WGAN)[12] Number of epochs=4000		
Improved Training of Wasserstein GANs (WGAN_GP)[13] Number of epochs=30000		

Figure 115: Results produced from various type of GANs