Universidad de Alcalá

PhD. Program in Electronics: Advanced Electronic Systems. Intelligent Systems

Anomaly detection for video surveillance applications

PhD. Thesis Presented BY

Mohammad Ibrahim Sarker

Advisors

Dra. Marta Marrón Romera

Dra. Cristina Losada Gutiérrez

Alcalá de Henares, January 15nd, 2024

Resumen

Esta Tesis Doctoral presenta un enfoque innovador para mejorar la seguridad pública mediante la automatización de la detección de anomalías en secuencias de videovigilancia, abordando amenazas críticas como robos, accidentes y comportamientos antisociales. En el centro de esta investigación está la integración del Aprendizaje de Múltiples Instancias (*Multiple Instance Learning*, MIL), una metodología de aprendizaje supervisado débilmente, con arquitecturas de aprendizaje profundo (*Deep Learning*, DL) de vanguardia. Esta fusión representa un avance notable en el campo de la videovigilancia y la detección de anomalías, aprovechando eficazmente las fortalezas combinadas de ambos: MIL y técnicas avanzadas de visión por computador.

El estudio introduce tres arquitecturas principales de DL: Attention 3D-ResNet-152, Transformer 3D-ResNet-152 y modelos de Transformer 3D-ResNet-152 Ensemble. Todos ellos están meticulosamente diseñados para detectar y clasificar eventos anómalos en escenarios del mundo real, mostrando un rendimiento excepcional en conjuntos de datos extensos como ShanghaiTech y UCF-Crime. El modelo Attention 3D-ResNet-152 sobresale en la extracción de características espaciotemporales, mientras que el Transformer 3D-ResNet-152 aprovecha las redes de Transformers para capturar dependencias de largo alcance en datos de video. El modelo Ensemble 3D-ResNet-152 Transformer, que combina las fortalezas de los anteriores, ofrece una robustez y precisión sin igual.

Los hallazgos de este estudio subrayan la eficacia del sistema de detección de anomalías propuesto, demostrando su capacidad para aprender de datos débilmente etiquetados y detectar anomalías con precisión. La combinación de MIL con arquitecturas de DL de última generación, incluidos los mecanismos de atención y Transformers, representan un avance importante en la tecnología de videovigilancia.

Una contribución significativa de esta investigación es el desarrollo de una arquitectura de detección de anomalías en tiempo real, que integra los modelos avanzados de DL con capacidades eficientes de procesamiento en tiempo real. Este sistema, respaldado por una interfaz gráfica de usuario (*Graphical User Interface*, GUI) amigable, permite un procesamiento rápido de los *frames* de video y una detección de anomalías oportuna, mejorando la aplicación práctica del sistema propuesto.

Con todo ello puedo concluir que esta investigación no solo proporciona una solución robusta, eficiente y escalable para la detección de anomalías en escenas de videovigilancia

y en tiempo real, sino que también establece un nuevo punto de referencia en el ámbito de la seguridad pública, allanando el camino para futuras innovaciones en estas tecnologías de vigilancia.

Palabras clave: Detección de Anomalías, Análisis de Video de Vigilancia, Arquitecturas de Aprendizaje Profundo (DL), Aprendizaje de Múltiples Instancias (MIL), Tecnología de Seguridad Pública, Procesamiento de Video en Tiempo Real, Visión por Computador, Aprendizaje Supervisado Débilmente.

Abstract

This doctoral thesis introduces a novel method to improve public safety through the automation of anomaly identification in surveillance video sequences. It specifically focuses on identifying and mitigating significant risks such as robberies, accidents, and antisocial conduct. This research focuses on combining Multiple Instance Learning (MIL), a form of weakly supervised learning, with state of the art Deep Learning (DL) architectures. This fusion represents a significant progress in the fields of video surveillance and anomaly detection, skillfully utilizing the combined capabilities of both MIL and modern computer vision techniques.

The study commences with the initial contribution, which centers on the application of Video Anomaly Detection (VAD) for detecting anomalies in video-surveillance scenarios. This study introduces a weakly supervised learning technique that classifies video clips into normal and abnormal occurrences. The algorithm utilizes spatio-temporal data obtained from a pre-trained temporal convolutional 3D neural network, known as T-C3D. The proposed approach provides a novel ranking loss function to minimize the occurrence of false negatives and improve the differentiation between regular and anomalous videos.

The second contribution enhances the first one by adding a real-time anomaly detection system that is poorly supervised and adds an attention mechanism. This method enhances the process of extracting features and exhibits exceptional performance and efficiency in real-time applications, extensively evaluated on benchmark datasets.

After conducting these fundamental investigations, the dissertation presents three main DL architectures: Attention 3D-ResNet-152, Transformer 3D-ResNet-152, and Ensemble 3D-ResNet-152 Transformer models. These models are intricately crafted to identify and categorize unusual events in real-life situations, demonstrating outstanding performance on vast datasets such as ShanghaiTech and UCF-Crime. The Attention 3D-ResNet-152 model is highly effective at extracting spatio-temporal features, whereas the Transformer 3D-ResNet-152 model utilizes transformer networks to capture long-range dependencies in video data. The Ensemble 3D-ResNet-152 Transformer model, which integrates the characteristics of many models, provides exceptional resilience and precision.

The findings from this study underscore the efficacy of the proposed VAD system, demonstrating its capability to learn from weakly labeled data and accurately detect anomalies. The combination of VAD with state of the art DL architectures, including at-

tention mechanisms and transformers, signifies a major advancement in video surveillance technology.

A significant contribution of this research is the development of a real-time VAD architecture, which integrates the advanced DL models with efficient real-time processing capabilities. This system, supported by a user-friendly Graphical User Interface (GUI), enables swift processing of video frames and timely anomaly detection, enhancing the practical application of the proposed system.

Putting everything in a nutshell, this research not only provides a robust, efficient, and scalable solution for real-time VAD but also sets a new benchmark in the realm of public safety, paving the way for future innovations in surveillance technology.

Keywords: Video Anomaly Detection (VAD), Surveillance Video Analysis, Deep Learning (DL) Architectures, Multiple Instance Learning (MIL), Public Safety Technology, Real-Time Video Processing, Computer Vision, Weakly Supervised Learning

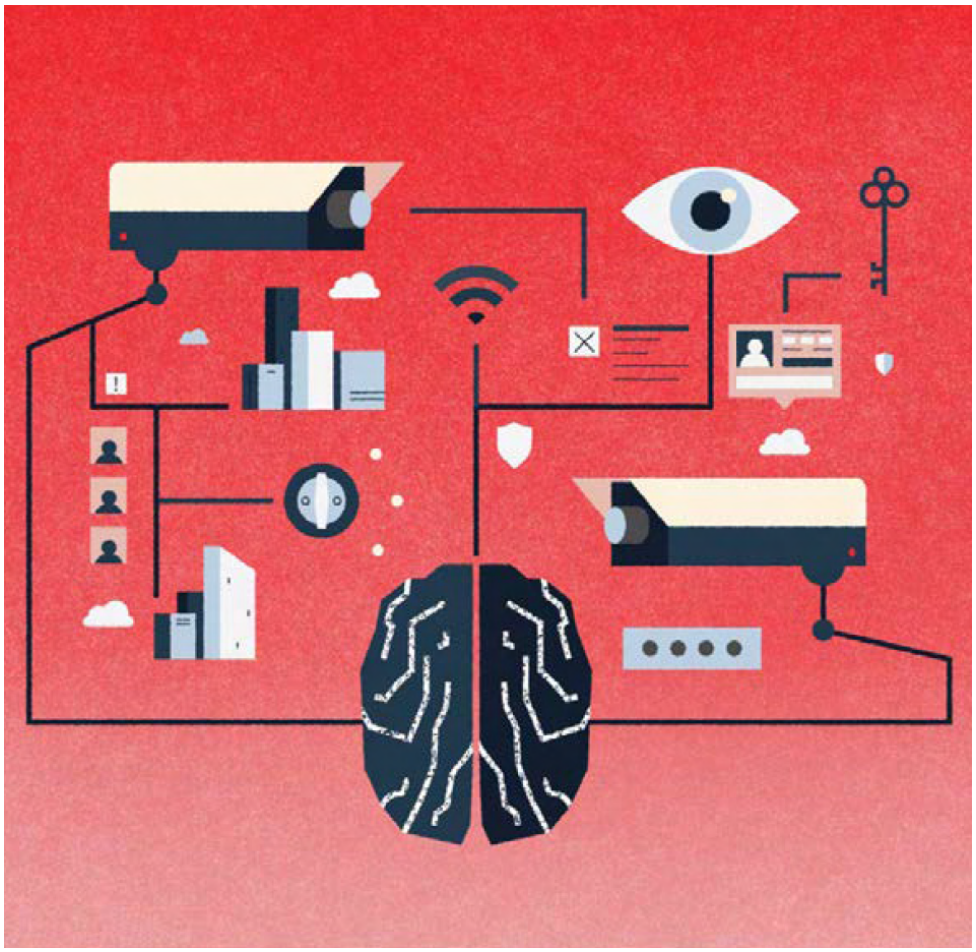
Contents

List of Figures

List of Tables

Chapter 1

Introduction



This Doctor of Philosophy (PhD) presents a proposal for Video Anomaly Detection (VAD) based on Multiple Instance Learning (MIL), a weakly-supervised learning method, that is explored in the thesis in this context. Moreover, within the MIL architecture, feature extraction is also used with different attention mechanisms, whose utility is deeply analyzed. Furthermore, for the purpose of completing the anomaly detection method, an improved ranking loss function is also presented.

Using features obtained with Temporal Convolutional 3D Neural Networks (T-C3D) [1], Inflated 3D ConvNet (I3D) [2] and 3D-ResNet-152 [3] within the attention mechanism context mentioned, allows for obtaining results in the pursuit objective that outperform the current state of the art on all best-known benchmarks in the area of VAD, according to experimental results.

Finally, two benchmark datasets for VAD, ShanghaiTech [4] and UCF-Crime [5], have been used to validate these proposals. In addition, a realistic in-the-wild dataset, named GBA [6] was recorded and made public for the scientific community, in order to complete the validation of the PhD proposal as a real engineering solution.

These results demonstrate that the attention mechanism proposed enhances the richness of feature extraction from videos for the normality classification task, so important in the objective of obtaining a global model that tackles the diversity and variability changes in the topic real-world scenarios.

All such contributions are complemented with real-time implementation, which has been stated as crucial in VAD context, as a sake of personal security, and appears as a challenging purpose in this Artificial Intelligence (AI) context.

Objectives, starting hypotheses and specific contributions of the PhD listed above are contextualized, and deeply analyzed in the following sections within this chapter.

1.1 Context

The introduction of surveillance cameras for the purpose of maintaining public safety has made a substantial impact in recent times, as it has effectively reduced security risks such as thefts, accidents, and unlawful behaviors. A challenging investigation topic thus appears: it is the VAD, in the big area of human behaviour analysis and scene understanding. By maintaining continuous surveillance of highly populated establishments such as banks, marketplaces, train stations, and bus terminals, although it cannot guarantee public safety, there is a possibility of reducing crime in these areas [7,8]. However, simply recognizing anomalies in real time is insufficient to prevent the loss of human life and financial resources. It is crucial to accurately identify these abnormalities and take suitable measures to effectively mitigate the associated risks.

In the present era, human operators primarily rely on manual detection to identify anomalies in surveillance videos. Hence, the efficacy of anomaly detection is heavily contingent upon the individual responsible for it. The individual must possess proper training and demonstrate unwavering dedication to the task, as it requires a significant investment of time and involves repetitive and monotonous activities. Consequently, the labor becomes tedious and repetitious because there are few varied and uncommon anomalous events, resulting in a significant likelihood of human error in detecting such events. Automation is necessary to enhance the dependability of security systems.

Several modern surveillance systems operate in a semi-automatic mode, requiring human intervention to classify and even detect anomalies after the technology has merely detected some basic actions there. This is because anomaly detection and categorization algorithms are not very reliable in real-world situations. Furthermore, recent methods [5,9] deal separately with the tasks of anomaly detection and categorization. Methods like the violent flow descriptor [10] or the heuristic approach [11] were used to distinguish between violent and non-violent videos.

Fortunately, well-known computer vision techniques are ultimately applied to video surveillance to detect anomalies. These algorithms can be created to spot unusual occurrences in camera photographs [12] and videos and provide details about the time period in which they took place. First and foremost, these algorithms ought to be able to identify unusual events in a surveillance video. Additionally, these algorithms can be improved to include such events in the normal category in the case the video is mistakenly identified as abnormal. Therefore, even if the occurrences are different from what would be seen in a surveillance video situation, these algorithms have to be able to discover and classify them correctly, whether they are connected to robberies, accidents, burglaries, violence [10], or aberrant actions [11].

However, because such methods are optimized to find a certain kind of anomaly, their capacity is still constrained.

The limited progress in anomaly classification methods can be attributed to the substantial enhancement in anomaly detection methodologies. In order to fully automate the current surveillance systems, it is important to develop a comprehensive and universal model for detecting anomalies in video scene interpretation. This model must possess the capability to accurately detect anomalies and categorize them in order to provide timely and suitable actions. This research proposes a model that aims to simultaneously train the detection and categorization of anomaly instances in real-world surveillance film. This is the main contribution in the field of Video Anomaly Detection VAD topic.

As mentioned, anomalous occurrences in real-life situations are not easily understandable, as they usually occur unexpectedly and might be of various kinds. Consequently, it is extremely difficult to develop a detector that consistently operates reliably in all video anomaly circumstances.

On the other hand, supervised procedures can be used when instances of both normal and abnormal occurrences can be recognized and classified. Compared to unsupervised algorithms, these strategies often offer a greater degree of precision in Video Anomaly Detection VAD. However, the process is time-consuming and demanding as the algorithm must undergo training for each video segment and be provided with a sufficient number of video instances depicting both normal and abnormal events for accurate detection. Moreover, these approaches can only detect the abnormal events that are included in the training.

In this context, unsupervised or weakly-supervised detection methods are favored due to their ability to identify an event even without prior knowledge or training. Both methods depend on the reconstruction error as a basis for classification during the testing phase. This error is calculated after constructing a training model using typical instances throughout the training process. Furthermore, the unsupervised technique has a drawback in that it may sometimes identify typical situations as anomalous due to the potential for routine events to evolve over time. This might result in a significant number of false positives.

The scientific community is thus researching weakly-supervised classification algorithms to address these problems, since they have a high degree of accuracy and can detect anomalies in video surveillance with little training.

A variety of techniques for Human Action Recognition (HAR) in videos has emerged as Deep Learning (DL) has flourished. In a stream network, the video is normally split into two sections, a spatial and a temporal one [13]. Information from the video portion is contained in the spatial part, where Residual Networks (ResNet) [14] have been proven to be highly effective for 2D Convolutional Neural Network (CNN) structures. Indeed, ResNet is a neural network built on the concept of skip connections that alleviate the degradation issue of accuracy saturation.

The work presented by Wang and Cherian in [15] introduces an innovative approach to anomaly detection, termed as Generalized One-class Discriminative Subspaces (GODS). This method, while distinct, can be conceptually likened to the advancements in video processing techniques. Notably, the I3D approach, discussed in [2], presents a similar paradigm shift in its domain. I3D, a significant technique in video processing for feature extraction, finds notable applications in VAD.

In this context, the Inception network's 2D convolutional filters are expanded to 3D, utilizing the I3D methodology. This adaptation, while offering improved performance in video analysis, comes with the trade-off of increased computational cost, as highlighted in [16]. Similarly, the GODS approach by Wang and Cherian establishes a new frontier in anomaly detection, pushing the boundaries of conventional methods, albeit in a different field and application.

1.2 Objectives

This PhD research is anchored on the objective of innovating an advanced method for enhancing anomaly detection in video surveillance data, that is VAD, by applying different supervision techniques within the classification task, as the MIL paradigm. The comprehensive objective is to unravel the unfolding incidents in a scene, primarily when it carries safety or critical significance. This includes monitoring activities and behaviors of people, as well as detecting anomalous events concerning objects [17].

The scope of anomaly detection in surveillance and security is vast and challenging. It encompasses tasks such as the detection of physical objects in complicated real-world scenarios, understanding human behaviors, and predicting anomalous events related to objects, all of which contribute to the anomaly detection framework [17].

Furthermore, the scientific community has broadly classified anomalies based on their frequency of occurrence, distinctive attributes from their normality model, or explicit meaning based on an abnormality model [18]. Each classification requires a unique detection technique, primarily relying on learning mechanisms of normality or abnormality. This research contemplates all alternatives and considers varying degrees of supervision based on the requirements of such complex application.

The general objective can be divided in several specific ones that are listed below:

1. Study, develop and implement a robust system for **extracting characteristics and descriptors from video surveillance scenes**.
2. **Develop a robust, real-time anomaly detection system for in-the-wild scenarios, using semi-supervised** models that leverage both physical and semantic characteristics previously obtained. This system is designed to efficiently detect anomalies in real time, building upon the robust and responsive architecture discussed earlier.
3. **Collect and collaborate in generating a multisensorial in-the-wild database,** named GBA [6], which is useful for the anomaly detection task.
4. **Exhaustively evaluate the proposals** on well-known datasets in VAD within the scientific community, as well as on those contributed in the PhD.
5. **Make the models developed available to the scientific community** through impactful publications and common software repositories for this purpose.

Overall, the objective of this research is to design a potent and efficient method for VAD. By integrating MIL, weak supervision, high-level feature extraction techniques, and optimized training strategies, the approach exhibits immense potential for real-world applications.

In order to tackle such objectives, in the following sections, firstly working hypotheses and then proposed contributions of the PhD are clearly stated.

1.3 Problem Hypotheses

Some hypotheses have been set in the research beginning stages, mainly tacking into account the related state of the art.

Trajectory-based VAD is an alternative used in conventional approaches [19]. The basic idea behind these techniques is that an observation video will be labeled as an anomaly if the objects of interest are not moving along the previously learnt normal trajectories. These techniques have rigorous monitoring requirements for the objects of interest, making them potentially unsuitable for in-the-wild footage. Additionally, when applied to a different domain, these conventional methods are ineffective, since they are unable to adapt to brand-new anomalies that they have never encountered before.

Some other traditional approaches, machine learning based, followed the same idea as the one presented, but in the same line, all of them lack the generalization needed for VAD already mentioned [5].

As DL algorithms have demonstrated a lot of success in the computer vision field, subsequent research in VAD has focused on the use of DL approaches. Figure 1.1 illustrates the performance comparison between DL based techniques and conventional approaches, such as traditional machine learning or basic neural networks.

As it can be seen in this figure, DL based techniques outperform these conventional approaches as data amounts increase. The inherent capabilities of DL models, including their ability to capture complex patterns and adaptively learn from large datasets, contribute to their superior performance. Thus, DL techniques have achieved remarkable success in domains like computer vision, natural language processing and speech recognition, and, of course VAD, revolutionizing tasks such as image classification, object detection, machine translation, and speech synthesis.

Therefore we decided to conduct our study to reinforce the effectiveness of DL based algorithms in surpassing traditional methods, with the ability to extract valuable insights from data, and driving advancements in AI.

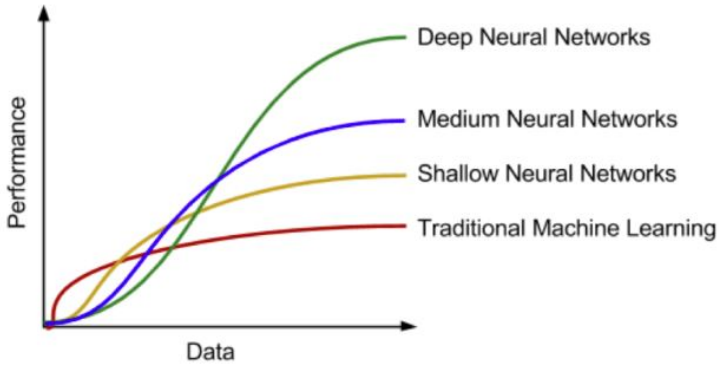


Figure 1.1: Performance comparison of DL methods and traditional ones [20] in AI.

Since VAD is actually a task within video processing, there are two kinds of characteristics that may be extracted by DL algorithms: the spatial features and the temporal ones.

In this context, CNN [21], one of the best-known DL models now in use, have shown to learn high-quality visual features or representations. Moreover, during the past ten years, a big number of techniques have been proposed for extracting temporal features, and the majority of them have obvious benefits and drawbacks. For instance, by including a transmission mechanism within hidden layers, the fundamental Recurrent Neural Network (RNN) [22] meets the goal of retrieving temporal information. Unfortunately, long-term videos are still inaccessible for RNNs.

Therefore, the design of a model for VAD that can overcome the drawbacks of previous methods tackling the spatio-temporal features in long and diverse sequences in-the-wild scenarios is still a challenging and novelty research topic selected to be the other PhD starting hypothesis.

1.4 Contributions

As previously mentioned, the primary aim of this study is to design and develop a system for VAD, based on MIL, a variant of supervised learning also referred to as weakly-supervised learning. The proposal aims to improve the effectiveness of VAD while enhancing its generalization capability. Thus, the key contributions to this dissertation are as follows:

1. **Design a weakly-supervised approach for anomaly detection:** an improved framework for anomaly detection in video surveillance scenes is proposed. This framework leverages the MIL paradigm, enabling the model to recognize local patterns and anomalies at different time scales and providing a learning setup to distinguish deviations from typical behaviors [23].

   An initial version of the proposed framework has been published in an international journal indexed in the Journal Citation Reports™ (JCR) [23]: "***Semi-Supervised Anomaly Detection in Video-Surveillance Scenes in the Wild***". 2021. Sensors 21, no. 12: 3993. (https://doi.org/10.3390/s21123993).

2. **Introduction of an improved ranking loss function with an Attention 3D-ResNet-152 model, a Transformer [24] 3D-ResNet-152, and an ensemble 3D-ResNet-152 Transformer:** we demonstrate that these models, in conjunction with a newly designed ranking loss function, facilitate a comprehensive representation of video content by distinguishing between typical and anomalous behaviors. This method amplifies the model's discriminative capacity, thereby improving its ability to differentiate between normal and anomalous events [14].

   A second approach including the integration of an attention mechanism has been published in the proceedings of 2023 European Symposium on Computer and Com-

munications (ESCC 2023) [25]: "***Real-time Weakly supervised Anomaly Detection with Attention Mechanism***".

3. **Comparison of different state of the art CNNs for feature extraction:** an exhaustive evaluation of the attention 3D-ResNet-152 model, the Transformer 3D-ResNet-152 one, and the Ensemble 3D-ResNet-152 Transformer model was conducted. This comparative study revealed each model's strengths and limitations, which guided the selection process for the most efficient feature extraction model [2, 24] in our VAD architecture posterior application.

4. **Real-time application of the proposed model:** the proposed model was tested in various real-time scenarios, highlighting its practical applicability and potential for real-world deployment. Moreover, a simple Graphical User Interface (GUI) was also developed to test the proposal in real-time scenarios, as a PhD colophon.

5. **Performance analysis on well-known datasets:** the proposed method was tested against state of the art weakly-supervised VAD datasets. Superior performance was demonstrated on large-scale datasets, including the well-know UCF-Crime [5], and ShanghaiTech [4]. Additionally, it was tested with the our GBA dataset, for validating the proposal in-the-wild scenarios, while utilizing features from the UCF-Crime dataset without using any single video from GBA dataset for training, achieving impressive results.

It is worth mentioning that in addition to working in the contributions described above, which has led to the publication of a journal and a conference paper, the work carried out during the development of the PhD has also allows collaborating with other researches in related topics, allowing the preparation of a journal paper related to human action recognition [26], as well as a conference paper on anomaly detection [27] that are listed below:

- "***3DFCNN: Real-time action recognition using 3D deep neural networks with raw depth information***". Sánchez-Caballero, Adrián; de López-Diz, Sergio; Fuentes-Jimenez, David; Losada-Gutiérrez, Cristina; Marrón-Romera, Marta; Casillas-Pérez, David and **Sarker, Mohammad Ibrahim**. Multimedia Tools and Applications. 81, 24119–24143. 2022. https://doi.org/10.1007/s11042-022-12091-z.

- "***A Proposal on Stampede Detection in Real Environments***". Cob-Parro, Antonio Carlos; Losada-Gutiérrez, Cristina; Marrón-Romera, Marta; Gardel-Vicente, Alfredo; Bravo-Muñoz, Ignacion; **Sarker, Mohammad Ibrahim**. Proceedings of the Eleventh International Conference on Indoor Positioning and Indoor Navigation, Lloret de Mar, Spain, 29th November, 2021. http://ceur-ws.org/Vol-3097/paper29.pdf.

1.5 Proposed Solution

To address VAD challenges analyzed in previous sections, we introduce an improved method. Figure 1.2 illustrates a general diagram of the proposed system. As seen, the anomaly detection proposal includes several stages, briefly described below and explained in detail in chapter 6.

First (left in pink and light blue in the figure), the proposal incorporates a module for preparing the weak supervision and temporal segmentation on video sequences, in order to optimise and generalize the anomaly identification process. The algorithm has to be trained to recognise local patterns and anomalies at various time scales using these digestible pieces from the sequences, as in [28].

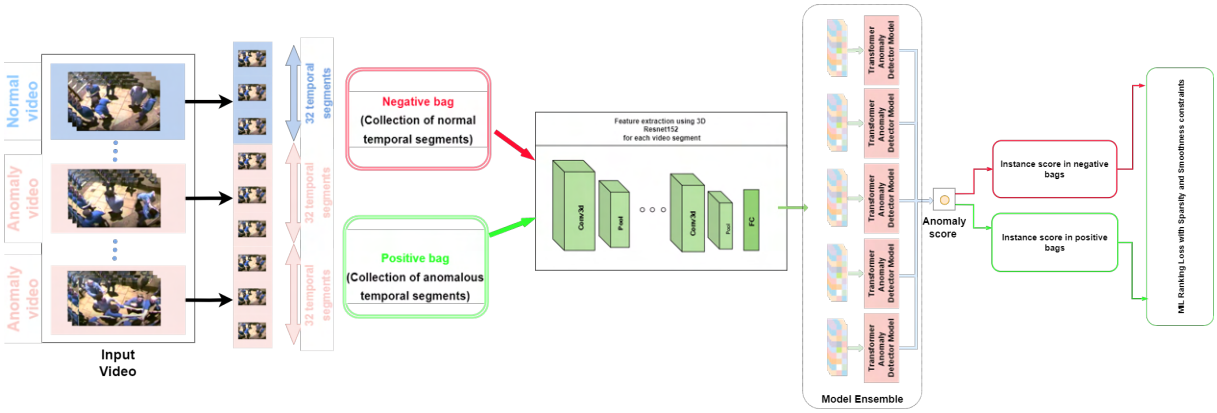


Figure 1.2: Final proposal for anomaly detection in surveillance videos.

The segmented video data is then used to create positive and negative bags in accordance with the MIL paradigm [29] (in green and red at the left of figure 1.2). Positive bags signify there aberrant activity, while negative ones customary behaviour. The objective is to provide the model with a contrasting learning setup So that it can more easily recognize deviations from typical behaviors, in order to be adaptable to non-trained anomalous scenes that appear in real-world applications. Therefore, the proposed strategy tries to maximise anomaly detection effectiveness while reducing dependency on manual annotation.

The suggested system has tested to use different CNNs, analyzing its behaviour in the application of interest (3D-ResNet-152 based Attention and Transformer models), for feature extraction (central blocks in green in figure 1.2). By differentiating between regular and anomalous actions, these models offer a thorough description of video footage. These models for semi-supervised anomaly identification in video surveillance were subjected to a critical evaluation. In order to choose the most effective model, the comparison will be concentrated on the advantages and disadvantages of each model.

This research will include a multi-head self-attention mechanism from the Transformer architecture for feature significance determination and attention methods, to better op-

timise the semi-supervised anomaly detection model [24] (at the right part of figure 1.2, in colours). This tactic keeps in searching to increase the model's ability to discriminate between typical and atypical events by boosting its discriminating strength.

Last but not least, the model is also effectively trained using a negative ranking loss function. By enabling multi-class classification, these functions improve the model's performance in tasks as challenging as anomaly detection, and promise more precise and accurate outcomes (at the very right part of figure 1.2 in green and red again).

In conclusion, the suggested analysis and method proposed not only improves VAD accuracy but also lessens the demand for manual annotation and intervention. Along the PhD we show the method's ability to considerably advance the state of the art in automated video surveillance and VAD.

1.6 Outline

This PhD thesis is structured into eight chapters, each delving deeply into the topics of video surveillance and anomaly detection using deep learning techniques.

After this introduction, chapter 2, discusses the related work done by the other researchers on the area of interest, related to the different contributions of the PhD. Then, in chapter 3, the focus shifts to the datasets that form a core part of the research. This chapter details their sources and the preprocessing steps necessary for achieving the aims of the PhD study.

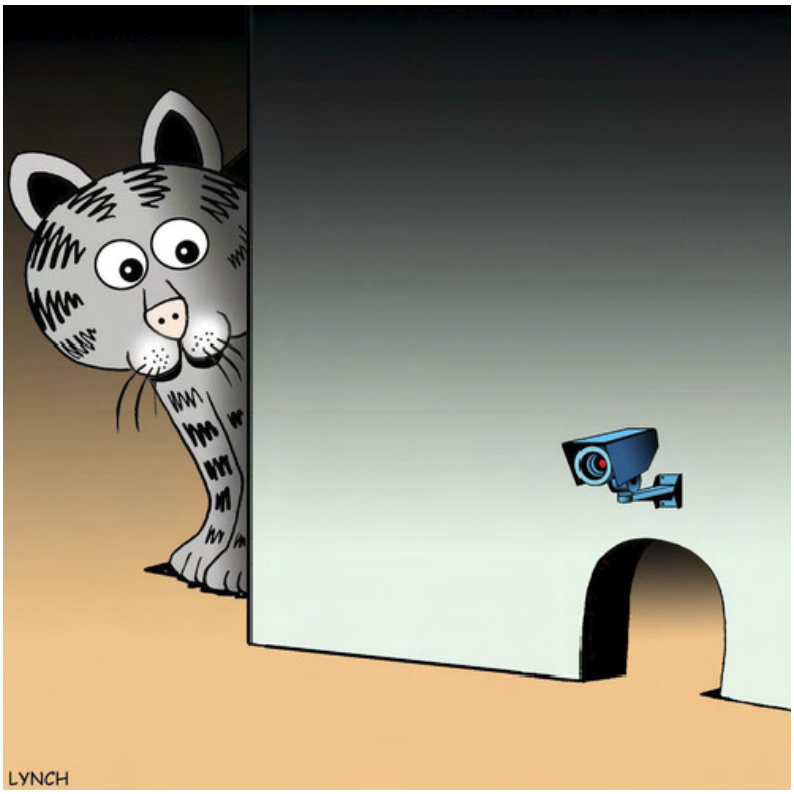
In chapter 4, we explore the utilization of weakly labeled training videos to identify abnormal human behaviors in surveillance footage, establishing the foundation for this research. Chapter 5 then introduces the integration of an attention mechanism into our study. This significant enhancement bolsters the anomaly detection process, with a particular focus on its efficiency in real-time settings.

Chapter 6 delves into the methodologies employed in this thesis, covering advanced architectures that have been harnessed to capture intricate spatial and temporal patterns in video data. In this chapter, we also discuss the development of a GUI for the Real-time Anomaly Detection Prototype and its testing in real-world scenarios.

Finally, chapter 6.8 presents the outcomes of this research. It offers a comparative analysis of the results obtained against existing benchmarks and discusses these findings within the broader context of the DL field, whereas chapter 7 summarizes the main conclusions and future works.

Chapter 2

Background



In this chapter we discuss many studies related to abnormal human behavior detection in video sequences. First, the main topic of interest is analyzed, i.e. the evolution of video surveillance, including the different approaches developed, focused on anomaly detection in real time applications and the use of attention mechanism in such context. We also review the previous works on anomaly detection in real time applications and the use of attention mechanisms.

After that, we analyze the different proposals of the scientific community for anomaly detection in video surveillance scenes based on a weakly supervised learning algorithm. There is also discussed the challenges of VAD and the use of MIL, in the state of the art of video in anomaly detection. Finally, it is explained where the proposals in this PhD are

different from other from the mentioned state of the art, elaborating where more research is still needed, and how to tackle it.

2.1 The Evolution of Video Surveillance

Video surveillance technology saw its early development during the V-2 rocket program at the end of World War II. Scientists, including the German Rocket Team and experts from the Applied Physics Laboratory at Johns Hopkins University, adapted the V-2 rockets for scientific purposes. They replaced the warhead section with scientific instruments, including movie cameras, to capture high-altitude images. This modification led to the first photograph of the Earth from space, taken on October 24, 1946, using a DeVry 35 mm black-and-white movie camera, which was traditionally used for newsreels. This pioneering use of movie cameras in rockets marked an early instance of employing video technology for surveillance and observation from a high altitude [30].

The first real security cameras were installed in the latter part of the 1960s and early 1970s. For instance, the police department in New York installed cameras on Time Square and in the Municipal Building to decrease criminal activity. Even still, crime rates stayed the same because these early devices could not record video for subsequent analysis, needing ongoing live monitoring [30].

After the Video Cassette Recorder (VCR) became widely available and inexpensive due to mass manufacture, security cameras began to be widely used. When cameras were installed in important London Underground stations in 1975, and used by store owners and banks to record evidence of robberies or other criminal acts, it became clear that this was the case. Nevertheless, VCR and analogue technology had considerable drawbacks, including frequent tape changes, wear and tear, and poor performance in low light. However, developments in electronics, cameras and surveillance technology in the 1990s, including the Charge-Coupled Device (CCD) chip, digital multiplexing, and motion-only recording, reduced some of these drawbacks [31].

The devastating World Trade Centre attack on September 11, 2001, prompted a greater focus on surveillance video processing, leading to the surge in popularity of real-time video surveillance for human detection. Facial recognition technology has been used since then, in airports for identity verification, schools to find missing children and sexual offenders, and at the Statue of Liberty site to stop terrorist assaults [32], just talking the United State of America use, due to terrorism fear.

Video Energy Vector (VEV) [33] marks a significant advancement in real-time anomaly detection in surveillance video processing. They introduced the VEV to efficiently reduce feature map dimensions while preserving essential spatio-temporal information. Using a Support Vector Machines (SVM), they enhanced the system's ability to identify and classify various abnormal events. This development is particularly pertinent in the United

States, where enhanced surveillance is crucial for public safety and counter-terrorism efforts.

Due to improved computing power and internet accessibility, video surveillance has evolved exponentially in recent years. More than one camera can be connected now to just one computer for analyzing their sequences, and video data can be stored for longer periods of time thanks to improved compression algorithms and higher storage capacity.

In such context, automatic VAD is evidenced to be a trending topic since the last two decades, still not successfully solved, and thus tackled in this PhD.

2.1.1 The Rise of Surveillance Cameras

The increasing number of cameras is leading to a greater accessibility and the affordability of video surveillance equipment, driven by the escalating concerns regarding crime and security threats. While it is not currently mandatory to register cameras for this purpose, it is undeniable that the number of installations is on the rise.

More specifically, and according to New York Community Board District (NYCLU), the number of cameras increased by more than five times, from 769 to 4468, between 1998 and 2005 [34]. This is an increase of around 580% in just seven years.

Moreover, as reported by [35], Spain has a total of 1.96 million security cameras, which translates to approximately one camera for every 24 individuals. There has been a substantial shift in the country's strategy towards security and surveillance, as indicated by the noticeable rise in the deployment of surveillance cameras.

The number is even greater in Denmark, according to [36], with an estimated number of 350.000 cameras, or one for every 17 people. These devices are extensively utilized, indicating a global pattern of increasing surveillance for various objectives, such as public safety, security, and crime prevention.

Deeper debate about the cybersecurity of such devices could be tackled but is well far from this PhD, but the high number of video surveillance systems again reveals the interest of the application here focused.

2.1.2 Societal Implications of Video Surveillance

A multitude of studies have been conducted on the effects of video surveillance, with a primary focus on its role in crime deterrence. An insightful review was recently presented by the Spanish Ministry of Interior [37]. In their publication, "Crime prevention effects of closed circuit television: a systematic review" [37], 22 evaluations of various studies were conducted, leading to the following observations:

- 11 studies found a positive effect on crime deterrence.

- 5 studies found no significant effect on crime.
- 6 studies found a negative effect on crime deterrence.

The diverse results of these research can be related to multiple variables. The effectiveness of surveillance cameras in deterring crime is dependent on the particular context and location in which they are used. Furthermore, the existence of surveillance cameras might occasionally result in the relocation of illegal activities rather than their decrease, as wrongdoers move their operations to regions with less monitoring. The use of surveillance technologies must strike a balance between security needs and safeguarding individual private rights, which are also influenced by concerns about privacy. In regions with more stringent privacy legislation, the implementation and usage of surveillance cameras may be more limited, which could potentially affect their efficiency.

This study's general finding is that using video monitoring does not significantly lower crime rates. Despite this, there are more surveillance cameras installed every day, and making the best use of them is still a top priority.

Hence, it can be inferred that the primary objections raised by society against video monitoring are infringements on privacy and concerns regarding the proliferation of a surveillance industry. Given the circumstances, individuals conveyed that they would be less apprehensive about potential breaches of their privacy if an automated system could be implemented, wherein computers solely process the video footage captured by these cameras. This constitutes the primary aim of the PhD.

2.1.3 Applications of Video Surveillance

Beyond the detection of unusual activity, video surveillance has many other uses. In Marbella, Spain, cameras with cutting-edge AI software are employed to comb through hours of video, in order to find suspects based on distinctive facial features, clothing colour, age, form, gender, and hair colour [38]. In addition, these cameras can recognise any vehicle's model and colour without needing to know its licence plate [38]. In figure 2.1, the capabilities of the new video surveillance system are demonstrated [38].

In a similar manner, cameras and post-processing software in Aalborg, Denmark, record license plates at various points to determine traffic flow [39]. In fact, in terms of monitoring criminal behavior, surveillance footage can be used either prospectively as proof, or proactively to identify crimes, mishaps, or other occurrences of interest [40].

In any case, the research highlights that there is still a very big open opportunity to AI for VAD, and thus to the PhD aim, as systems are getting more and more data, thanks to the mentioned increase of cameras and security interest by the society.

Figure 2.1: Local police in Marbella explaining its video surveillance system [38].

2.1.4 Future of Intelligent Video Surveillance

Active usage of video surveillance requires continuous observation to spot anomalous activities, which is frequently achieved by having numerous displays being watched by a human operator. Although humans are excellent situation interpreters, their effectiveness declines with more screens and longer monitoring periods. Moreover, the demand for human resources grows together with the number of deployed cameras. This poses a big financial hurdle because installing cameras is cheap but paying people to monitor the results is expensive [41].

Even if human ability to recognise anomalies may not be completely supplanted by automated abnormality detection in the near future, it has demonstrated to be a useful tool, and a step that has to be taken in the automation is pursued. Thus, the VAD objective of this PhD gets more importance.

2.2 What is Anomaly Detection?

Anomaly detection is an important data analysis task that identifies anomalous or abnormal instances within a given data. It is an interesting area of data mining research as it involves discovering enthralling and rare patterns in data.

Tan, in [42] defines anomaly detection as the task of detecting observations whose characteristics are significantly different from the rest of the data. Loy in [43] defines an anomaly as an event which has a low statistical representation in the training data. Chandola in [44] defines an anomaly as a pattern in the data that does not conform to a well-defined notion of normal behavior.

Although an anomaly is defined by researchers in various ways based on its application domain, one widely accepted definition is that of Hawkins [45]: "*An anomaly is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism*".

The subject of anomaly detection spans a wide range. For instance, noise detection processes the data to eliminate unnecessary noise so that the patterns in the data may be analyzed more thoroughly. Nevertheless, this is different from anomaly detection in that we are interested in locating those records rather than filtering them. Novelty detection is the identification of fresh patterns that were absent or ignored in the past. The primary difference between it and anomaly detection is that those patterns typically modify the standard, or already obtained, data model. The methods used to address those linked issues are frequently the same as those employed in outlier detection.

Uses of anomalies is very significant because of its wide range of application in people's life. It is widely used in medical and public health [46], fraud detection [47], network intrusion [48], video surveillance [49] etc. For example, an unusual pattern in communications traffic could mean that network domain is compromised and data has been hacked; anomalous behavior in credit card transactions could indicate fraudulent activities and an anomaly in a Magnetic Resonance Image (MRI) image may indicate the presence a tumor [50].

Therefore, in real world it is very hard to name all possible anomalous events because these events are very complicated and diverse. So the algorithms used in anomaly detection should not take into account previous anomalous events. As these methods serve to lessen human liability it is important that anomaly detection is completed with minimum supervision and maximum reliability.

In [51], as shown in figure 2.2, a simple introduction to the concept was found, in which normal events are modeled (spatially) by functions F1 and F2. Anomalies are assigned as A1, A2, A3, so modeling data is done in such a way that, from the mass of normal data, outliers are separated. This is a key factor in an effective anomaly detection system: an effective representation of events will decrease intra-class distance, and increase inter-class distance, such that events are closer clustered to similar ones, and further clustered from dissimilar ones. Poorly represented data (in training) shall be considered as an anomalous event, so these will be easily classified as outliers.

Anomalies may be also introduced into data for a number of causes, including malevolent action, such as credit card fraud, cyber-intrusion, terrorist activities, or system failure, but they all share the trait of being intriguing to the analyst. In fact and as a conclusion, the important aspect of anomaly identification is their *interestingness* or practical relevance of anomalies [44].

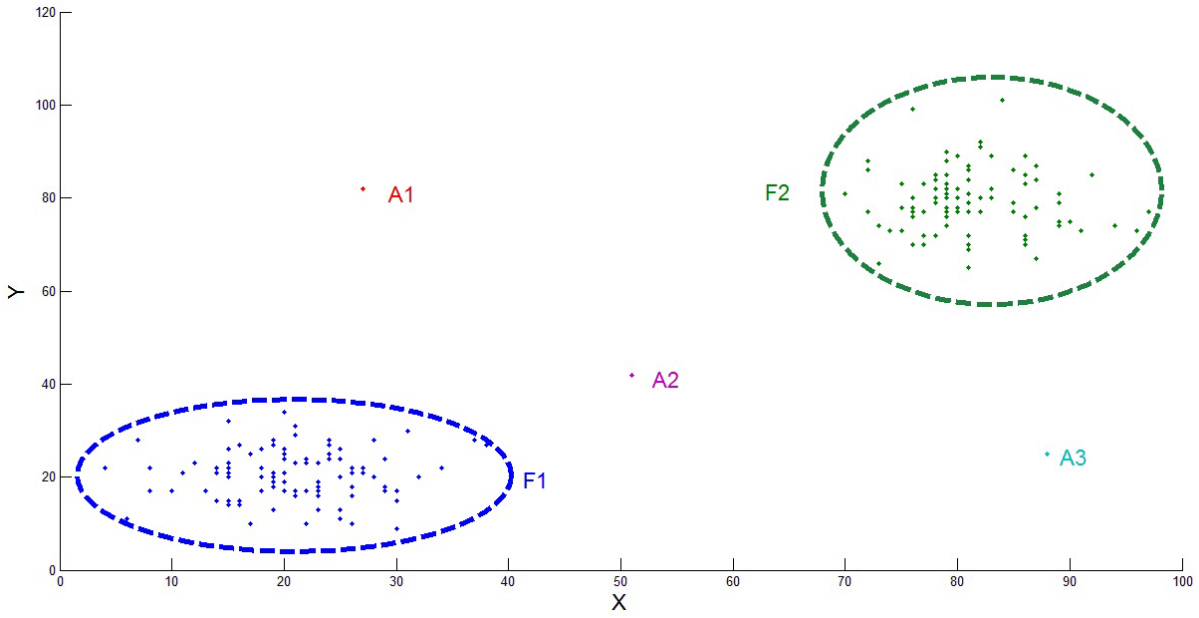


Figure 2.2: Simple example of anomalies in an arbitrary space. Events are modeled by functions F1 and F2, with anomalies being A1, A2, A3 [51].

2.2.1 Challenges

The anomaly detection problem is becoming more difficult to solve due to a number of issues. As analyzed previously, the distinction between normal and abnormal behavior is frequently arbitrary. However, the normal behavior in some fields, like intrusion detection, is continually changing, making those normal changes susceptible to being wrongly classified as outliers. Also, it is necessary to adapt the anomaly detection approaches to the various application domains. Additionally, the lack of labeled data for training and validation places restrictions on the findings and interpretations made.

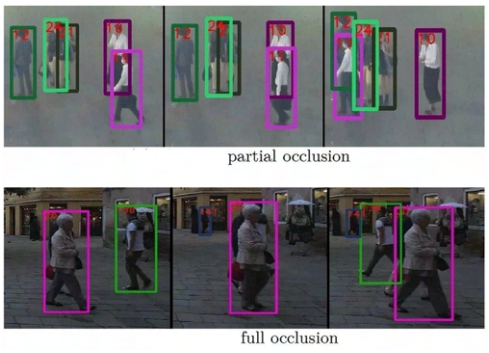
A straightforward approach to anomaly detection is defining a region in the behaviour model that represents normal behavior, and the remaining part that does not belong to the normal region declared as anomalies. But some factors make this approach very difficult [44]. Thus, challenges in VAD can thus be described as follows:

- Most existing methods for anomaly detection are supervised, where the data is available for anomalous events. But for real-world applications, it is not always possible to gather enough data for training such various and unexpected situations.
- The boundary between abnormal and normal behavior is not much accurate. It is very hard to clarify or make a region which includes all probable normal behavior. So, the anomalies that are very close to that margin in the model can be normal or abnormal.
- For different applications of anomalies, the viewpoint is different. For example, a small variation in normal observation might be diseases in the medical domain.

Whereas, in the stock market domain small variations might be considered normal. This makes it complex to apply a particular domain tested technique to another domain.

- Existing anomaly detection methods excel primarily in structured scenarios. For example, they are proficient at identifying anomalies such as traffic violations, like running a red light or exceeding the speed limit. However, in scenarios such as surveillance systems for public safety, anomalies may include behaviors like physical altercations or fights in crowded areas. These situations can be complex and dynamic, requiring specialized classifiers to accurately detect anomalies. Therefore, for each different type of problem, separate classifiers may be needed to train in order to detect anomalies effectively.
- Most VAD techniques use single event detection, so it fails to exploit the multi-perspective on an event.
- Also, very few anomaly detection techniques consider the interaction of multiple elements in an anomalous event. In a real scenario, there are many anomalies that are caused by joint elements, which can be considered as *collective anomalies*. For example, an accident on a highway can be caused by the behavior of both a driver and a pedestrian.
- Noise and occlusions are also challenging problems for VAD (see figure 2.3). Sometimes noise has a similar effect in the image than an anomaly, and it is difficult to distinguish between noise and an anomaly, because data containing noise is analogous to actual anomalies.

(a) Example of problem of noise.



(b) Example of an occlusion problem.

Figure 2.3: Examples of noise and occlusions problems in anomaly detection [52].

In our research, we address these challenges through a series of innovative methodologies. Our approach utilizes a combination of machine learning techniques that are adaptable to the dynamic nature of real-world scenarios. By leveraging both supervised and unsupervised learning models, we aim to create a more robust and versatile anomaly detection system suitable for a variety of applications. Our research particularly focuses

on enhancing the accuracy and efficiency of VAD in complex environments where traditional methods fall short.

2.2.2 Anomaly Types Depending on Application

Anomalies are data patterns that deviate from well-defined characteristics of usual patterns [52]. Furthermore, it has been noted that an essential aspect of anomaly identification is the inherent character or category of the deviation. There are three distinct categories of anomaly:

1. **Point anomalies:** the primary form of anomalies are point anomalies, which are singular occurrences of data that appear in distinct locations within the examined event space. These zones are commonly identified by their spatial distance from conventional data points. In simple terms, a point anomaly refers to a data point that deviates from the anticipated pattern of the dataset.

   In the specific context of the provided figure 2.4, Point 11 emerges as a notable anomaly. The dataset is seen to be increasing gradually and consistently in the picture, but at Point 11 the pattern is broken by an unusual spike. Fortunately, from Point 12 onward, the data quickly returns to its typical range.

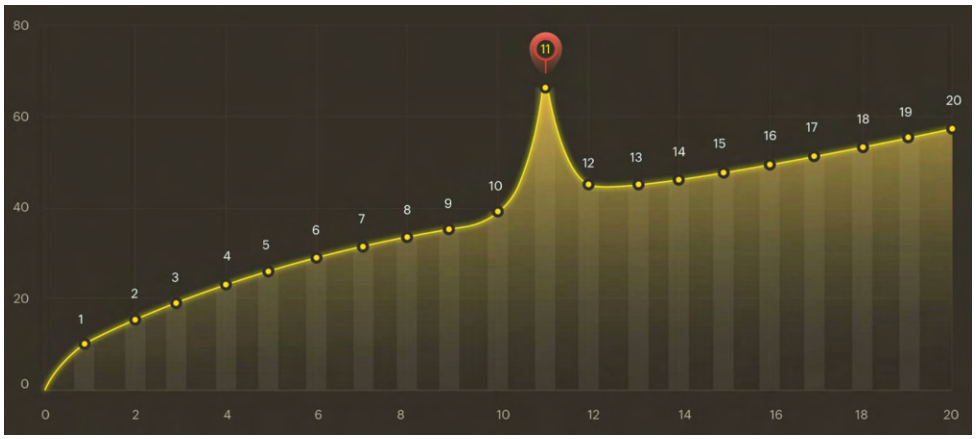


Figure 2.4: Example of point anomalies [53].

   Imagine a video surveillance system designed exclusively for monitoring a busy shopping mall. Typically, individuals exhibit consistent and foreseeable behavior, adhering to established patterns and routes. However, if an individual were to exhibit sudden odd behavior or engage in sprinting through the mall, it would be considered a point anomaly. The sudden deviation in the behavior of the video material from its usual pattern indicates the existence of an anomaly.

   As reported by Mehran, Oyama, and Shah [54] in their 2009 study, there was significant advancement in the field of abnormal crowd behavior detection using the social force model. Their research focused on identifying unusual behaviors in crowd

dynamics. This was achieved by applying the social force model, typically used in physics to study the motion of particles, to analyze human behavior in crowds. Their innovative approach allowed for the detection of abnormal crowd behaviors, which is a critical aspect in surveillance and security, particularly in areas with high foot traffic. This study has important implications for enhancing public safety measures through the effective monitoring of crowd behaviors.

In their paper, Marsden *et al.*[55] introduce a Holistic Features methodology for real-time crowd behaviour anomaly detection. This research is noteworthy due to its comprehensive methodology in assessing crowd dynamics, employing holistic characteristics that represent the complex nature of human actions in large gatherings. Their methodology is specifically built to function instantaneously, rendering it exceptionally pertinent for implementations in public safety and urban surveillance.

The research presented by Wang and Xu [56] delves into the domain of crowd behavior analysis, specifically targeting anomaly detection in large groups of people. The authors present a specialized real-time analysis technique called Spatio-temporal Texture Modelling, that focuses on modeling spatio-temporal textures. This technology demonstrates exceptional innovation by effectively analyzing intricate visual data in congested environments, enabling the detection of possible hazards or emergencies in heavily populated regions.

Another example at VAD is keeping an eye on a parking lot. The system possesses the capability to identify both incoming and exiting automobiles and has become familiar with the consistent parking behaviors. If an unfamiliar vehicle were to arrive and be parked in a location where it is not authorized, this could lead to a deviation from the norm. This detection is facilitated by comparing the observed behavior with established patterns of typical behavior.

Hence, the challenging aspect of anomaly detection in such cases lies in accurately modeling the space where the anomaly may be readily distinguished from normal patterns.

2. **Contextual anomalies:** each data instance has the two sets of attributes listed below that defines it:

   - **Contextual characteristics.** The contextual characteristics are utilized to establish the context (or vicinity) for the instance. Contextual features in geographic datasets include things like the longitude and latitude of a given site. Time, in the context of time series data, serves as a contextual factor that denotes the position of an event throughout the entire sequence.
   - **Behavioral attributes.** The behavioral attributes of an object determine its inherent qualities that are not influenced by the surrounding context. An instance of a behavioral characteristic in a dataset of spatial information is the

measurement of precipitation, which represents the average quantity of rainfall at every given location worldwide.

Hence, contextual or conditional anomalies manifest when data exhibits inappropriate behavior inside a certain circumstance. Anomalies play a crucial role in detecting suspicious occurrences in video surveillance. Due to heightened Christmas shopping, it is expected that a larger proportion of individuals would utilize credit cards for their transactions. Although this increased spending is atypical, the elevated use of credit cards during the Christmas period is deemed within the expected range, and so not considered abnormal. Conversely, the absence of any spending during a vacation period might be seen as an unusual occurrence within its specific context, as it deviates from the expected norm.

Let's consider figure 2.5, where a dip should be visible at regular intervals of 5. An unexpected abnormality is seen when the image is examined more closely: Point 15's predicted decrease is missing. It is common to assume that these drops have justifiable causes, yet this belief is unfounded. This departure from the expected pattern is a contextual abnormality.

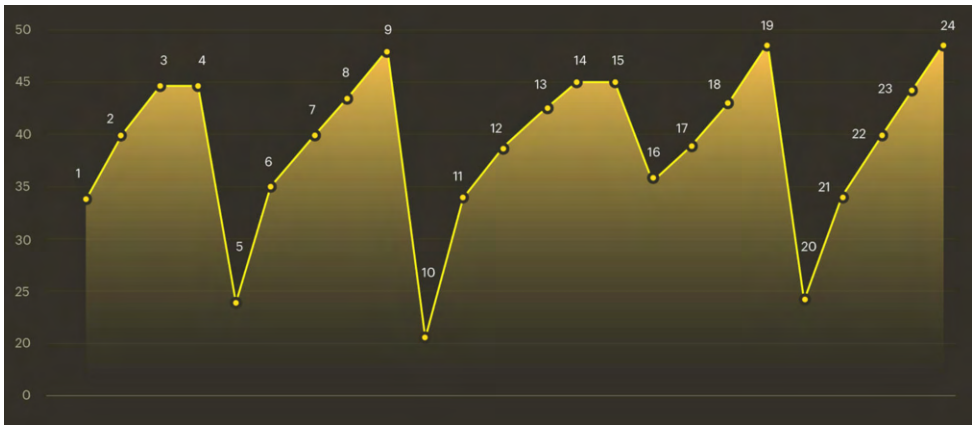


Figure 2.5: Example of contextual anomalies [53].

Thus, it is important to consider time in VAD because it is a key factor in anomaly identification. An incident's timing, for instance, may have an impact on how a scene is categorised in video surveillance: while some behaviours could be entirely normal during the day they may be aberrant during the night, and vice versa. By taking into account the temporal context, video surveillance systems can effectively recognise and categorise irregularities based on their time.

3. **Collective Anomalies:** instead of a single data event, an anomaly is considered to be collective when it is noticed within a sequence of data events.

   Imagine that during normal business hours, a video surveillance system is monitoring the movement of people in four different areas of a shopping mall. Each accounting is identified by a different colour in the surveillance data: orange, blue, green, and yellow in figure 2.6.

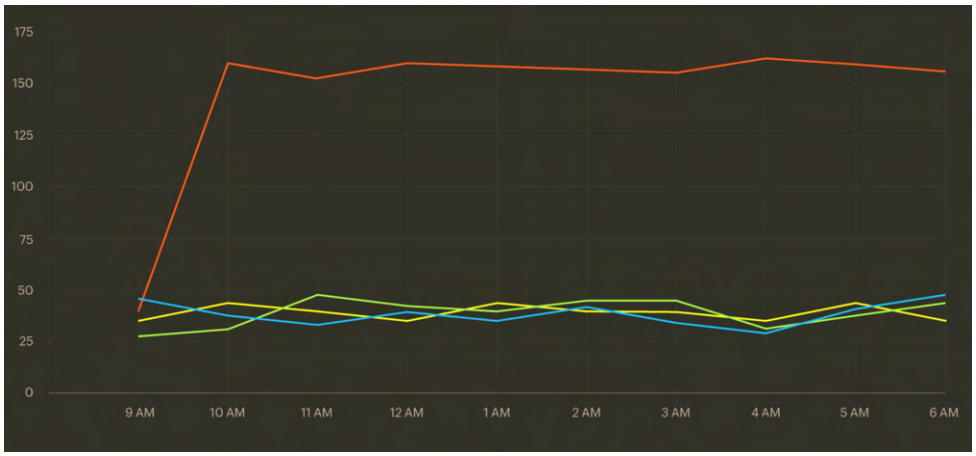


Figure 2.6: Example of collective anomalies [53].

The considerable foot traffic could initially appear normal if you only pay attention to the orange region. Nevertheless it is noticed to be peculiarly high when compared to the blue, green, and yellow accounting. If we had simply concentrated on the orange data, this chaotic pattern might have gone unnoticed. However, when data are seen in the context of the overall activity it is simple to identify the abnormality.

This example highlights how crucial it is to include collective data in video surveillance in order to accurately detect anomalies [4], i.e. add diversity in the Machine Learning (ML) process that is trained to detect them.

2.2.3 Techniques and Approaches

In this context, it is not necessary to classify every event into either 'Abnormal' or 'Normal' and often the events are left with an outlier score: a continuous variable representing the degree of separation within the set of normal behaviors. The results can then be better presented as a ranked list or visualized as a heat map of anomaly probabilities. Basically, the anomaly detection setup depends on the labels available in the dataset used for training, distinguishing between three main types.

In the following paragraphs an explanation about each of them, as well as its advantages and drawbacks for the anomaly detection task of interest are included.

1. **Supervised Anomaly Detection:** refers to a method of modeling in which an algorithm learns a model from a representative set of data, known as training data, in which each element is labeled with its true class membership [57]. While training, the algorithm establishes the relationship between data elements and the corresponding class membership. After the training is finished, if it is correctly performed, the algorithm has to be able to determine the class membership of new data elements not present in the training dataset. However, this method has some downsides. First problem is the amount of normal events with respect to abnormal ones that

is generally in the dataset used for training. Because of fewer number of abnormal events it is very difficult to obtain the representative set of examples needed to make a correct training. This phenomenon causes the results of anomaly detection to be falsified. In addition, the lesser volume of training events causes the system to provide results correctly or miss an event entirely. The most common supervised algorithms are: SVM, K-Nearest Neighbors (KNN), Bayesian Networks and Decision Trees [58].

2. **Unsupervised Anomaly Detection:** Among the three main types of anomaly detection, this method is the most flexible. In supervised anomaly detection, the same datasets are used for training and testing, but the difference is that the labels are not used for training. In contrast, in unsupervised anomaly detection, there is no need for labeled data during training. Instead, the algorithm scores for unsupervised anomaly detection based on the intrinsic properties of the unique dataset. In this method, class labels are thus imposed automatically by the technique, so we save the reliance on the labeled training data. Therefore, there is no reliance on the labeled training data. Abnormal events can be recognized by their uniqueness because it is presumed that when using these in unsupervised anomaly detection techniques, the frequency of normal behavior is significantly greater than that of abnormal ones, solving the downside of supervised techniques in such contexts. Additionally, abnormal events should be very different from the normal ones by some distinct characteristic known by the algorithm. Thus, overfitting in such characteristics may become a problem for unsupervised methods. The most common unsupervised algorithms are K-Means, Self-Organizing Maps (SOM), C-means, Expectation-Maximization (EM) meta-algorithm, Adaptive Resonance Theory (ART), Unsupervised Niche Clustering, and One-Class SVM.

3. **Semi-Supervised Anomaly Detection:** semi-supervised approaches either use a mixture of labelled and unlabeled data or use weakly labelled data [59]. Often some of either the normal or the abnormal events are labelled instructing the system how to segment the data into normal and abnormal classes. The weakly supervised methods normally use a large amount of automatically labelled data with known error, which does not overwhelm the training data [44]. Such methods are applicable when class labels can be estimated approximately, as hand labelling is very expensive.

In this thesis, we propose a weakly supervised learning algorithm whose functionality and objective is related to the semi-supervised approach previously described, as justified and stated below, in order to tackle one of the objectives stated in 1.2.

2.3 Weakly Supervised Learning

As it has been explained before, supervised learning techniques in ML construct predictive models by learning from a large number of training examples, where each training

example has a label indicating its Ground Truth (GT) output [60]. Although it has had superior success over other learning methods, it also has its difficulties. The high cost of data labeling is associated with the achievement of strong supervision through supervised methods. So for DL or ML, weakly supervised methods are recommended for its cost-benefit ratio.

Moreover, MIL is a form of weakly supervised learning where training instances are arranged in sets, called bags, and a label is provided for the entire bag. As discussed in the previous subsection on challenges in VAD, this method addresses specific difficulties encountered in anomaly detection, such as the scarcity of labeled data and the complexity of defining what constitutes an anomaly. For these advantages in addressing the outlined challenges in the VAD application of interest, such ML techniques are analyzed in detail in this thesis, in the following sections.

2.3.1 Types

Typically, three methods are associated with weakly supervised learning. For a better understanding, an example is shown in figure 2.7 and explained below.

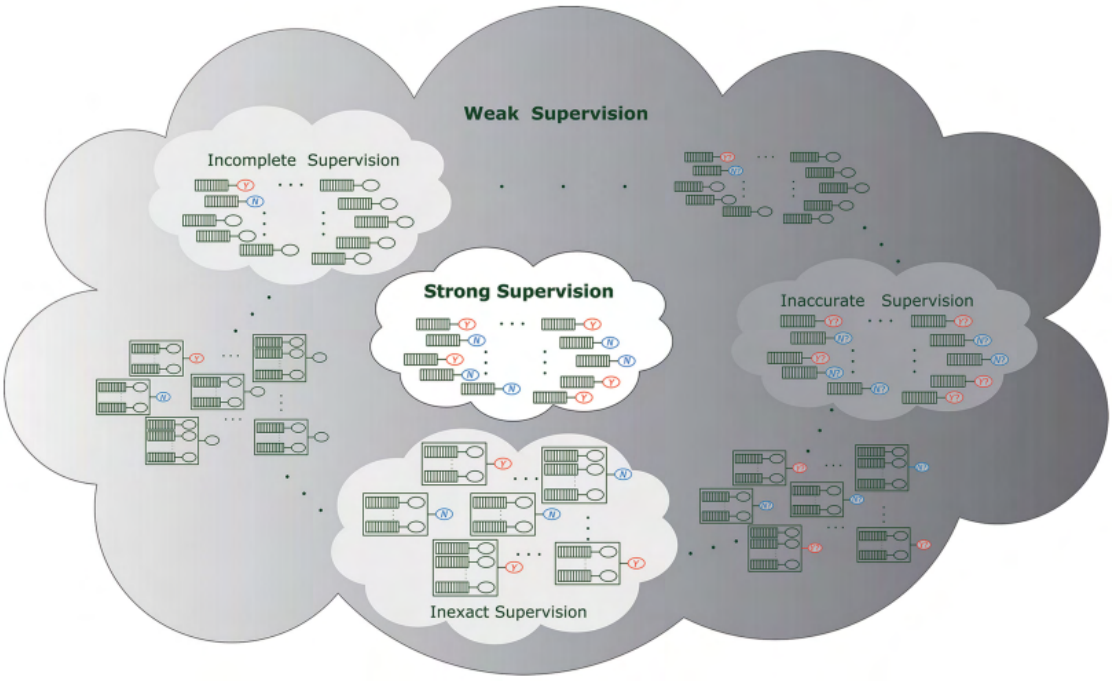


Figure 2.7: Illustration of three typical types of weak supervision in cloud graphs. Bars denote feature vectors; red('Y')/blue('N') marks their labels; '?' ('Y?'/'N?') implies that the label may be inaccurate. Intermediate graphs depict some situations with mixed types of weak supervision [60].

1. **Incomplete Supervision:** in incomplete supervision among all data only few labels are given. This situation is common for different tasks. For example, human annotators give GT labels in image categorization: although it is possible to obtain large quantity of images from the internet, due to human cost just a small subset of

images can be annotated. Formally the task in such type of supervision is to learn the classification function from a training dataset with a number of unlabeled instances. The rest of conditions of this weakly supervision are the same of a supervised learner with strong supervision [60].

2. **Inexact Supervision:** to understand inexact supervision lets consider the image categorization task again. To successfully implement it, it is desirable that all objects in the image are labeled. But only coarse-grained labels are possible to be provided [61]. This type of supervision is inexact supervision. The MIL is one of the major approaches of Inexact Supervision.

3. **Inaccurate Supervision:** the third type is inaccurate supervision, i.e., the weakly classifier trained with data whose labels are not always the correct GT. Such situation occurs, e.g., in VAD when the image annotator is careless or weary, or some images are difficult to be categorized [60], being a typical scenario learning with noisy labels [62]. Most theoretical studies assume random classification noise, but in reality, identifying the potentially mislabeled data and then correcting it is the idea of this weakly supervision. For example, in [63] such supervision is used in order to perform a preliminary filtering, that improves the classifier accuracy as a result of minimizing the noise problem to a certain extension.

   Crowd-sourcing is commonly used as a cost-saving way to collect labels for training data in ML. In [64] it is presented a review of how inaccurate supervision occurs with crowd-sourcing. In such context inaccurate supervision can be understood like this: by using different algorithms for the same task and modelling the task difficulty for weighting their results, it can be expected that better performance can be achieved. Thus, some approaches try to construct probabilistic models and then adopt an EM algorithm for the estimation [65]. In [66] they are presented some other learning examples based on a this inexact weak teacher (or crowd labels), closely related to learning with noisy labels.

Finally, it is crucial to acknowledge that, while we can categorize weak supervision into three distinct types - incomplete, inexact, and inaccurate supervision - in practical scenarios, these types often intermingle. This overlap is especially common in extensive datasets where ensuring perfect, detailed labeling for each instance is impractical, if not impossible. For instance, large-scale data collection, like crowd-sourcing, inherently involves a mix of inexact and inaccurate supervision due to human error and the impracticality of labeling every data point precisely. Thus, it's typical in real-world applications to encounter a blend of these supervision types, necessitating sophisticated machine learning methods that can adapt to and manage these complexities effectively.

2.3.2 Multiple Instance Learning (MIL)

Multiple Instance Learning (MIL) is an extension of semi-supervised classification, where training class labels are assigned to collections of patterns, known as bags, rather than individual patterns. Although each pattern is supposed to have an associated real label, it is understood that pattern labels can only be indirectly accessed through the label linked to its bags [67].

The MIL setting has numerous notable applications. Examples of applications include the categorization of molecules in pharmaceutical design, the indexing of images for content retrieval, the extraction of information from papers, and the identification of land mines. Each of these apps exhibits the same form of label ambiguity. However, the challenge with this learning technique lies in identifying the instances that decide the classification of bags as either positive or negative in the context of anomaly detection.

For binary classification, if at least one of the set is positive it can be easily termed as a positive sample. Because of MIL's characteristics to set the pattern of classifier, its main challenge is to distinguish the samples, whether the bags are actually positive or not. Thus, the goal of MIL is to classify unseen bags of instances based on the labeled bags as training data. Some variants of MIL [61, 68] also estimate the instance labels during the test phase, which would imply temporal localization of the symptom instances within the time segments.

Using the reference of [68] the problem definition of MIL can be described as follows: consider a set of feature vectors $x_i/i = 1, 2, \cdots, N$. Under a supervised learning framework, each feature vector i.e. instance would have a label $l_i = f(x_i) \in \{+1, -1\}$. In contrast, a weakly supervised MIL framework does not include labels for each instance, rather, it has labels for a set of instances, i.e. a bag. In other words, if the observed bag label is positive then at least one of the bag's instances must have produced that positive result. Furthermore, if the observed label is negative, then none of the bag's instances could have produced a positive result.

Dietterich *et al.* in [61] proposed the first solution to the MIL problem. It involved finding an axis-parallel hyper-rectangle in the classification space that captures the concept. In [69] two new formulations of MIL, as a maximum margin problem, are presented. Moreover even SVM was proposed as a learning method to solve MIL (named SVM for MIL), involving a mathematical optimization problem which can be solved by a mixed integer quadratic program.

Finally, it is worth to highlight that more recently Sultani *et al.* in [5] proposed a method to learn anomaly in VAD by deep MIL. In this method all surveillance videos (normal and abnormal) were divided into short clips, named bags. During training for anomaly detection, long untrimmed videos are divided into small segments (i.e. bags) and

fed into the network. This proposal gives a strong starting point for the contributions presented in this PhD.

2.3.3 MIL Applications

MIL approach has found diverse applications in various fields, including the military, where it has proven to be a powerful tool for solving complex classification, regression, ranking, and clustering problems, as stated below.

1. **Classification:** bag and instance classification are the 2 types of tasks that can be tackled by MIL. Instance classification is different from bag one because while training is performed using data arranged in sets in both of them, the objective is to classify instance individually just in the first one. In the work [67] it has been observed that the loss function in the learning process for the two types cannot be the same. However, for bag classification, at bag level it is not very problematic to misclassify an instance. For example, if some true negative instance are classified as positive, the bag label would not change. So loss function mostly depends on the structure of the problem. As a result, the performance of an algorithm for bag classification is not representative of the performance that could be obtained with it for instance classification. Moreover, many methods proposed for bag classification do not reason in instance space, and thus, often cannot perform instance classification. MIL classification is not limited to assigning a single label to instances or bags. The bags need to be labeled in multiple, keeping in mind that they can contain different concepts because of different instances in the bag.

2. **Regression:** solution for MIL regression has been approached in many ways. This task focuses on assigning real value to a bag. Some methods assign the bag label based on a single of its instances. This instance may be the closest to the target aimed [70], or the best fit in the regression model [71]. Another use of MIL in regression, replacing bag level classifiers by a regressor, instead of using a weighted combination of the instances in a bag to represent it as a single feature vector.

3. **Ranking:** one of the most active research area in ML is learning to rank. Instead of individual scores, these techniques main focus is improving relative scores. In [72], for improvement retrieval quality of search engines rank, SVM was introduced. In [73] the authors proposed an algorithm for solving multiple instance ranking problems using successive linear programming and demonstrated its application in hydrogen abstraction problem in computational chemistry. Recently, deep ranking networks have been used in several computer vision applications and have shown state of the art performances. In fact, there has been multiple more practical applications of MIL in ranking images. For example, feature learning [74], Graphics

Interchange Format (GIF) generation [67], face detection and verification [75], person re-identification [76], place recognition [77], data leakage prevention [78], image retrieval [79], etc.

There are some approaches in which rather than focusing on obtaining the real value, the objective of these methods is to compare the scores for sorting.[73], is proposals from the scientific community where ranking is performed by bag and instance respectively within the MIL context.

4. **Clustering:** current works in the scientific literature on the topic of clustering is very small. In this task clusters need to be found on a set of unlabeled bags. Clustering can be performed even at the instance level with MIL. For example, in the work [80], the algorithm identifies the most relevant instance of each bag, and performs maximum margin clustering on its instances. In [81] clustering is performed in the bag space using standard MIL algorithms and set-based distance measures.

The task pursued in this PhD matches with the first application, where MIL has demonstrated hopeful results in the state of the art, as commented previously and developed after within this work.

2.4 State of the Art in Anomaly and Video Anomaly Detection

Once the topic of interest is well reviewed and analyzed, we give a thorough overview of several anomaly detection techniques in the PhD area and go over their key features. Based on their supervised (Sup.) or unsupervised (Unsup.) nature, the goal they seek, the architecture they use, and the year of publication, these methods are contrasted.

2.4.1 Anomaly Detection

AutoEncoder (AE) has been intensively used in last decade in order to build reconstruction models that help in the task of anomaly detection in diverse applications.

Before delving into the specifics of VAD, it is important to understand the wider landscape of DL techniques in video analysis. The comprehensive survey presented in [3] offers valuable insights into DL methods for video captioning, which share similarities with anomaly detection techniques. This context sets the stage for a deeper exploration of specific DL-based anomaly detection methods detailed below.

2.4.1.1 Reconstruction

As a first example of a proposal from the scientific community, the Diverse Density framework [82], can be pointed out as a semi-supervised technique that uses AE with Mean Square Error (MSE) as the loss function for reconstruction-based anomaly identification.

Reconstruction-based anomaly detection has emerged as a compelling approach in unsupervised learning, primarily due to its ability to identify deviations from normal data patterns through the process of data reconstruction. This methodology is intriguing as it is predicated on the notion that while normal data can be reconstructed accurately, anomalies (being irregularities) result in higher reconstruction errors.

The effectiveness of reconstruction-based anomaly detection lies in its foundational principle: anomalies, inherently deviating from the normative data patterns, manifest in increased reconstruction errors. This discrepancy is a quantifiable metric for anomaly detection, pivotal in domains demanding outlier identification for safety and operational integrity. Its applications span network security [83], financial fraud detection [84], healthcare monitoring [85], and industrial systems [86] among others [44].

Notable techniques in this area include: the Replicator, introduced by Hawkins *et al.* in [87], which utilizes neural networks for data replication; RandNet, proposed by Chen *et al.* in [88], that employs an ensemble of AE; and Recurrent Discriminative Autoencoder (RDA) by Zhou *et al.* in [89] and Unsupervised sequential Outlier detection with Deep Architecture (UODA) [90] by Mabrouk *et al.*, that again based on AE and RNN, make use of MSE as loss function.

These methods demonstrate the versatility and efficacy of reconstruction-based techniques in detecting anomalies across various application domains.

2.4.1.2 Generative and Predictive

In the same line, in recent years there has been a significant expansion of anomaly detection in generative and predictive models. This burgeoning interest is driven by the promise of innovative techniques that offer advanced capabilities for identifying and understanding anomalies across various domains.

For instance, the Adversarial Neural Network for Anomaly Detection (ANOGAN) method [91] combines CNNs and generative models, utilizing the Mean Absolute Error (MAE) loss function, to effectively detect anomalies in tomographies. Similarly, the Energy-Based Generative Adversarial Network (EBGAN) [92] approach employs a combination of a CNN, a Multi-Layer Perceptron (MLP) and a Generative Adversarial Networks (GAN) for semi-supervised anomaly detection, showcasing the versatility of generative models in VAD.

Moreover, predictive modeling plays a crucial role in anomaly detection, exemplified by techniques like Future Frame Prediction (FFP) [4] and Latent Space Autoregression (LSA) [93], both based on CNNs and utilizing loss functions such as MSE and a combination of MSE and Kullback-Leibler (KL) divergence. These advancements have practical applications in diverse fields, including event detection in visual telecommunications, where the ability to accurately predict and identify anomalies is vital for safeguarding data integrity and system security.

2.4.1.3 Classification

Another important group of works in anomaly detection can be found with classification objective: distance-based, one-class and, in general, clustering are different examples of such anomaly detection type.

In this context, there are several semi-supervised techniques. GeoTrans [94] and E^3 Outlier [95] use a CNN for event classification and Cross-Entropy (CE) as the loss function.

On the other hand, DeepOne-Class (OC) [76] focuses on one-class anomaly detection using a 3D CNN with the hinge loss, whereas DeepSemi-Supervised Anomaly Detection (SAD) [96] is a semi-supervised method that combines CNN and MLP networks for one-class anomaly detection, using the hinge loss, in different application fields, such as VAD.

Regarding unsupervised approaches, one notable example is the method of Random Distances Prediction (RDP), as described by Hu Wang *et al.* in [97]. This method employs MLPs for distance-based anomaly detection, using loss functions such as hinge loss and MSE. Advantages of this approach include its ability to learn complex data representations without labeled data, and drawbacks involve potential challenges in selecting appropriate distance metrics and thresholds for anomaly detection. This method is particularly relevant to this PhD, as it underscores the potential of MLPs in unsupervised learning scenarios, aligned with my focus on leveraging DL for anomaly detection.

Another method is the AutoEncoder based one-class Support Vector Machine (AE-1SVM), proposed by Minh-Nghia *et al.* in [98], that combines AEs and one-class SVM for anomaly detection. Advantages include the method's robustness in handling high-dimensional data and its effectiveness in identifying outliers. Drawbacks include the computational complexity of training and the consequent need for careful tuning of hyperparameters. The integration of AEs with SVM is of interest to the research in this PhD as it has demonstrated to be a hybrid approach to anomaly detection, combining the strengths of different learning paradigms.

Deep Embedded Clustering (DEC) [99] focuses on clustering tasks, using MLP networks with KL-divergence as loss function. Its main advantage lie in its efficient handling of large datasets and the ability to uncover hidden structures in data. However, a drawback appears on the quality of initial feature representations. This method's emphasis on clustering provides insights into alternative unsupervised learning strategies that could be adapted in the PhD.

Lastly, Deep Autoencoding Gaussian Mixture Model (DAGMM) [100] utilizes clustering with AE and a Gaussian Mixture Model (GMM) using a likelihood loss. Its strengths include the ability to model complex data distributions and detect subtle anomalies, whereas the main challenges of this approach includes again the need for extensive parameter tuning and potential consequent overfitting.

2.4.1.4 Scoring

Finally, it is worth mentioning the different anomaly scoring methods that have been designed by the scientific community in the last years in different applications.

Self-trained Deep Ordinal Regression (SDOR) [101] is an unsupervised method that focuses on anomaly scoring using ResNet and MLP networks with MSE as loss function in VAD applications. Abnormal Event detection based on an Heterogeneous Information Network Embedding (AEHE) [102] is another unsupervised method that combines an AE and a hierarchical ensemble for anomaly scoring, using softmax as loss function.

Analyzing weakly supervised methods, Sultani *et al.* [5] is based on 3DCNN and MLP networks, utilizing hinge loss. This approach has been deeply studied in this PhD due to its high reliability and applicability to datasets similar to those of interest.

Some other weakly supervised methods for anomaly scoring in VAD that may be highlighted due its reliability and architectures are: Deviation Networks (DevNet) [103], that employs MLP networks with standard deviation-based loss; and Adversarially Learned One-Class Classifier (ALOCC) [104], that combines an AE and a CNN with a GAN loss, leveraging adversarial training for anomaly detection that enhances robustness.

One-class Adversarial Nets (OCAN) [105] is also a semi-supervised technique focusing on anomaly scoring based on MLP networks and Long Short-Term Memory (LSTM)-based AE, leveraging GAN loss, where the use of LSTM networks appears to be especially suitable for time-series anomaly detection.

Finally Fence-GAN [106] and One-class novelty detection using GANs (OCGAN) [107] are other semi-supervised methods using CNN with GAN loss of interest. Fence-GAN creates a boundary in the data space that can be useful for further data analytics, while OCGAN excels in feature extraction and anomaly identification in image data. These methods highlight the effectiveness of adversarial approaches in VAD.

2.4.1.5 Summary

This section concludes with a thorough organization of the different anomaly detection techniques, among those presented, in order to get the best from each proposal. Table 2.1 presents the chronological spectrum of these anomaly detection techniques, spanning almost two decades.

These techniques can be understood in terms of their objectives, underlying architectures, supervision methods, and loss functions employed. In fact, anomaly detection techniques presented in previous paragraphs and in Table 2.1, vary in their objectives, as stated below:

- **Reconstruction-based** methods, such as RandNet [88], primarily use AE to recreate input data and identify deviations.

- **Generative models**, represented by ANOGAN [91] and EBGAN [92], aim to generate new samples, typically employing CNN.
- Techniques with a **predictive** objective, like FFP [4], use architectures like CNN to forecast future data points.
- **Classification-based** methods, such as GeoTrans [94], use CNN to label data as normal or anomalous.
- **One-class** techniques, exemplified by AE-1SVM [98], differentiate between data from a single class and anomalies.
- **Clustering-based** methods, such as DEC [99], cluster data and identify outliers.
- Lastly, there are models that determine **anomaly scores**, like SDOR [101], which assign a score based on the likelihood of a data point being an anomaly.

Moreover, as observed in Table 2.1, methods vary depending on whether they are supervised or unsupervised. Each approach has its own advantages and disadvantages, aligning with the analyses previously discussed in this chapter.

- While **supervised approaches** are reliable for classification tasks, they may struggle with anomaly detection in open settings, as justified in previous sections. Thus, semi-supervised ones, like GeoTrans [94] are more adequate here.
- **Unsupervised techniques**, such as DEC [99] and DAGMM [100], excel in clustering jobs but can be challenging to tune effectively.
- The benefits of both supervised and unsupervised methods are combined in **semi-supervised techniques** like OCAN [105] or Sultani *et al.* [5]. These methods enhance performance and adaptability by integrating a small amount of labeled data with a larger unlabeled dataset. Due to their improved accuracy and robustness in real-world situations, semi-supervised learning emerges as a promising approach for anomaly detection and also for VAD, which is of particular interest in this PhD. The most suitable method selection ultimately depends on the specific requirements of the application and the nature of the data, as stated here.

It is finally worthy to summarize the main weaknesses of these current techniques described, in the anomaly detection task:

- **Over-reliance on AEs**: while AEs, as seen in UODA [90], are effective at data reconstruction, they may struggle to spot subtle anomalies in high-dimensional data.
- **Under-utilization of 3D Convolutions**: the limited adoption of 3D convolutions, despite their capability to process spatio-temporal data, is evident, with only a few methods like DeepOC [76] employing them.

;

Table 2.1: Comparison of anomaly detection methods. Organized depending on their objective (application), underlying architectures, supervision method (unsupervised i.e. Unsup., semi-supervised i.e. Semi., and with weakly supervision i.e. Weak), and loss function employed.

Method		Objective	Supervision	Architecture	Loss	Y
Replicator	[87]	Reconstruction	Unsup.	AE	MSE	20
RandNet	[88]					20
RDA	[89]		Semi.	AE	MSE	20
ANOGAN	[91]	Generative	Semi.	CNN	MAE	20
EBGAN	[92]			CNN & MLP	GAN	20
FFP	[4]	Predictive	Semi.	CNN	MAE/MSE	20
LSA	[93]				MSE & KL	20
GeoTrans	[94]	Classification	Semi.	CNN	CE	20
E^3 Outlier	[95]					20
RDP	[97]	Distance	Unsup.	MLP	MSE	20
AE-1SVM	[98]	One-class	Unsup.	AE & CNN	Hinge	20
DeepOC	[76]		Semi.	Convolutional 3D Neural Networks (C3D)	Hinge	20
DeepSAD	[96]			CNN & MLP		20
DEC	[99]	Clustering	Unsup.	MLP	KL	20
DAGMM	[100]			AE & MLP	Likelihood	20
AEHE	[102]	Anomaly Scores	Unsup.	AE & MLP	Softmax	20
SDOR	[101]			ResNet & MLP	MAE	20
ALOCC	[104]		Semi.	AE & CNN	GAN	20
OCAN	[105]			LSTM-AE & MLP		20
Fence-GAN	[106]			CNN & MLP		20
OCGAN	[107]			CNN		20
Sultani *et al.*	[5]		Weak	C3D & MLP	Hinge	20
DevNet	[103]				Deviation	20

- **Stagnation in Architectural Innovation**: many models continue to rely on earlier structures, such as basic MLPs, potentially overlooking advancements in DL.
- **Inconsistency in Supervision Methods**: the varying supervision types, from unsupervised to semi-supervised, hint at the challenges tied to acquiring comprehensive labeled datasets for this task.

In conclusion, this section has provided a comprehensive overview of various anomaly detection techniques, considering their objectives, architectures, supervision methods, and loss functions. We observe a diverse range of strategies, from reconstruction-based and generative models to predictive, classification-based, one-class, and clustering-based approaches, each with its distinct focus and methodologies.

The methods detailed here align with the goals of our research, particularly in their use of sophisticated DL techniques for robust feature representation and anomaly scoring, stated as the first objectve in the PhD (see section 1.2). Our approach in this PhD

work shares similarities with these methods, especially in terms of leveraging advanced architectures and loss functions for effective anomaly detection in video surveillance.

Furthermore, the analysis underscores the need to balance the benefits of supervised, unsupervised, and semi-supervised approaches. Semi-supervised techniques, combining the strengths of supervised and unsupervised learning, emerge as particularly promising for anomaly detection in video surveillance, resonating with our research focus, as the second objective of the PhD proposes in section 1.2.

The identified weaknesses in current techniques, such as over-reliance on autoencoders, under-utilization of 3D convolutions, stagnation in architectural innovation, and inconsistency in supervision methods, guide us in developing our contributions. Our aim is to address these gaps, particularly in the context of VAD, by integrating the insights gained from this analysis into the design of our PhD contributions, and validate them in different contexts, stated as third objective in section 1.2.

This analysis is important in order to take it into account in the PhD contributions' design (in section 1.4), and will be translated to the more specific VAD task in the following section.

2.4.2 Video Anomaly Detection (VAD)

Rooted in the foundational concepts from chapter 1 and the methodologies outlined in chapter 6, this section provides an in-depth exploration of the complex and dynamic field of Video Anomaly Detection VAD. The incorporation of ML and DL techniques in Video Analytics for VAD plays an integral part in surveillance, protecting human safety and advancing the field.

The progression of VAD methodology commences with conventional supervised techniques and then advances towards more adaptable unsupervised and weakly supervised approaches. In the early stages of development, the SVM for MIL technique, as proposed in [69], marked a notable transition towards using machine learning for anomaly identification. The method employed support vector machines inside a framework of multiple instance learning, establishing a foundation for following techniques.

The utilization of non-parametric modeling, as demonstrated in the study [108], provided a novel viewpoint by emphasizing a more adaptable method for identifying anomalies in videos. This approach addressed the difficulties posed by pre-established models by adopting a non-parametric approach, hence enabling greater flexibility in detecting abnormalities.

The incorporation of holistic features in [55] was a significant advancement, as the approach utilized extensive characteristics of video data to identify anomalies, thereby enabling more sophisticated analysis. This method was crucial in comprehending the wider framework in which anomalies occur.

The technique of spatio-temporal texture modeling, as described in the paper [56], integrates spatial and temporal data to enhance the comprehension of video content with more accuracy and reliability. This approach was highly efficient in differentiating between typical and atypical patterns in video sequences.

As the field advanced, techniques such as Lu *et al.* [109] and Stacked-RNN [110] were developed. Sparse Combination Learning provided a distinct method for modeling regular behaviors, whereas the Stacked-RNN leveraged the capabilities of recurrent neural networks to capture temporal relationships in video data.

The study conducted in [111] on spatio-temporal AE and the research by [4] on FFP models highlight the growing trend in the field of utilizing deep learning techniques for anomaly identification. These approaches played a crucial role in acquiring intricate representations of video data and forecasting future states to detect abnormalities.

The GODS model, described in the publication [15], and the Incremental Spatio-temporal Learner, discussed in the publication [112], have made significant progress by emphasizing more dynamic and context-aware methodologies. The GODS model, specifically, was remarkable for its capacity to analyze intricate video data and identify irregularities in demanding settings.

The paper [113] introduced a novel approach called Mem-AE, whereas the paper [8] utilized both AE and LSTM models, demonstrating an advancement in the application of autoencoders. The Mem-AE model incorporated memory components into the autoencoder design, hence improving its capacity to store and identify regular patterns. Additionally, the fusion of AE and LSTM architectures capitalized on the respective advantages of both models, resulting in efficient anomaly identification.

The concept of utilizing restricted labeled data to train more effective models has been introduced by weakly supervised approaches, such as AE with limited supervision in Hassan *et al.* [114], Deep MIL in Sultani *et al.* [5], and Graph-based MIL in [115]. These approaches played an important part in diminishing the reliance on large annotated datasets.

Recent advancements, such as the Motion Aware technique [9], Zhang *et al.* [116], and Graph Convolutional label Noise cleaner (GCN) [117], have expanded the field by introducing motion-aware anomaly detection, temporal convolutional networks, and graph convolutional networks, with different features such as C3D or Temporal Segment Network (TSN). These methods exemplify the current state of VAD, demonstrating the incorporation of intricate neural network structures and sophisticated machine learning algorithms.

The framework proposed by [118] introduced a new approach by employing particle filtering for the purpose of anomaly identification. This technology was notable for its distinctive approach to monitoring and analyzing motion patterns in videos.

The recent advancements in the field of anomaly detection involve the integration of attention mechanisms with C3D. Notably, attention mechanism based MIL and T-

C3D have made significant progress in this regard. Therefore, ***Attention Mechanism based MIL*** [25] and ***Temporal Convolutional 3D Network*** [23] proposed in this PhD contribute in a remarkable shift towards more advanced and contextually aware algorithms in VAD.

Each of these approaches, ranging from the initial SVM for MIL to the latest developments in GCN and T-C3D, enhances our comprehension of anomaly identification in surveillance scenarios with more depth and subtlety. This evolution is not only consistent with the technological advancement mentioned in previous chapters, but it also emphasizes the difficulties and continuous advancements in VAD. A comprehensive functional comparison of many of these techniques, with regards of the ones here proposed, is shown finally in chapter 6.8, providing a clearer understanding of their efficacy and suitability in different monitoring situations.

The range of VAD methods, including unsupervised techniques like Spatio-temporal AE and Mem-AE, as well as weakly supervised methods like Deep-MIL and Attention Mechanism based MIL, demonstrates the dynamic progression of the field. The incorporation of these varied strategies highlights the intricacy of anomaly detection in surveillance and the significance of creating flexible models and data handling procedures, which is a key focus of this research as stated in earlier chapters.

Table 2.2 presents the chronological list of VAD techniques described in previous paragraphs, organized according the supervision level that apply. Moreover, datasets used to validate each of them are also inserted in the table. It has to be noted that chapter 3 is specifically dedicated to the analysis and definition (when needed) of datasets referred in Table 2.2, as most of them are used for validating the algorithmic proposals in this PhD.

2.4.3 Distinctiveness of the Proposed Methods

Once the global analysis of the state of the art is completed in this section, it is used in order to state the main distinctive characteristics of the PhD contributions.

These highlights are listed below and analyzed in detail in the following chapters:

- **Superior 3D Analysis**: Attention 3D-ResNet-152 and Transformer 3D-ResNet-152 harness 3D convolutions, offering a refined analysis of spatio-temporal data, which most mentioned methods miss.
- **Innovative Integration of Attention and Transformers**: the PhD proposal merges attention mechanisms to focus on significant data segments, and Transformers for contextual interpretations, reflecting a leap over many methods analyzed.
- **Ensemble Approach for Robustness**: combining the strengths of individual models, the ensemble of different DL models promises enhanced resilience and accuracy in VAD, presenting a formidable alternative to the techniques aforementioned.

Table 2.2: Comparison of VAD methods. The ones proposed in this PhD are marked in ***bold italics***.

Method	Year	Datasets
Unsupervised		
SVM for MIL [69]	2002	MUSK1, MUSK2
Non parametric modeling [108]	2012	UCSD-Ped1, UCSD-Ped2
Lu *et al.* [109]	2013	UCSD-Ped1, Subway, Avenue
Holistic Features [55]	2016	UMN, Violent-Flows
Spatio-temporal Texture Modelling [56]	2016	UMN, UCSD-Ped1, UCSD-Ped2
Stacked-RNN [110]	2017	Avenue, UCSD-Ped1, UCSD-Ped2, Subway
Spatio-temporal AE [111]	2017	Avenue, UCSD-Ped2, CUHK
FFP [4]	2018	Avenue, UCSD-Ped1, UCSD-Ped2, ShanghaiTech
GODS [15]	2019	JHMDB, UCF-Crime, Sonar and Delft pump
Incremental Spatio-temporal Learner [112]	2019	Avenue, UCSD-Ped1, UCSD-Ped2
Mem-AE [113]	2019	MNIST, CIFAR-10
AE and LSTM [8]	2019	Avenue, UCSD-Ped1, UCSD-Ped2
Video Event Completion (VEC) [119]	2020	UCSD-Ped2, Avenue, ShanghaiTech
VEV [33]	2020	UCF-Crime
Weakly Supervised		
Hassan *et al.* [114]	2016	Avenue, UCSD-Ped1, UCSD-Ped2, Subway
Sultani *et al.* [5]	2018	UCF-Crime
Graph-based MIL [115]	2018	Avenue, UCSD-Ped2
Motion Aware [9]	2019	UCF-Crime
Zhang *et al.* [116]	2019	UCF-Crime
GCN [117]	2019	ShanghaiTech, UCSD-Ped1, UCF-Crime
ParticleFiltering-based Framework [118]	2021	UCSD-Ped1, UCSD-Ped2, LIVE
Attention Mechanism based MIL [25]	2023	UCF-Crime, ShanghaiTech
Temporal Convolutional 3D Network [23]	2021	GBA, UCF-Crime, The Web-Dataset

2.5 Conclusions

This chapter has methodically explored the historical progression and current advancements in the field of anomaly detection, with a particular focus on VAD. The research has revealed the evolution from initial surveillance techniques to the intricate integration of ML and DL in modern VAD systems. This historical and technological overview is crucial for identifying the exact moments where this research might make a major contribution.

The examination of approaches, ranging from SVMs in MIL to the most recent advancements in DL techniques, discovers a field full of potential for creative advancements. Significantly, although current systems have achieved significant progress, they also present deficiencies and difficulties, especially in managing the intricacy and variety of surveillance data. This observation is vital, as it emphasizes the need for more versatile, effective, and situation-aware anomaly detection models in VAD.

The discourse surrounding different DL methodologies, such as AEs, generative models, and convolutional networks, underscores the wide range and promise of VAD investigation. Nevertheless, it also highlights the limitations of these methodologies, particularly in effectively analyzing spatio-temporal data and incorporating sophisticated attention mechanisms and Transformer architectures, which also contain attention blocks. These observations are crucial, as they steer the focus of my research towards addressing these particular deficiencies.

The complete evaluation indicates a significant requirement for creative solutions that improve the precision, flexibility, and contextual comprehension in VAD. This PhD aims to meet these demands by introducing innovative methodologies that utilize the advantages of sophisticated DL techniques while overcoming their existing constraints. The forthcoming sections of this thesis will reveal these contributions, illustrating how they not only address the highlighted deficiencies but also advance the field of VAD.

Chapter 3

Datasets for Anomaly Detection

Being HAR and anomaly detection in video surveillance applications prominent areas of research within the ML and DL domains, it is essential to consider that an action perceived as normal in one situation or application might be deemed abnormal in another. Consequently, a common approach in longitudinal studies, as stated in previous chapter, involves characterizing normal event patterns (regardless the ML supervision method used) and identifying abnormal events with any of these different ML approaches from these typical patterns [120].

3.1 Introduction

Several datasets have been extensively used to facilitate research in this area, and thus, will be used in this PhD, including the ones that follow, where the main reasons for their selection are also commented:

1. **UCF-Crime dataset** [5]: contains videos of various criminal activities, such as theft, vandalism, and assault. Researchers can use these videos to train and test models that detect and recognize criminal actions in real-world scenarios, converting the dataset in a baseline for comparing those models.

2. **ShanghaiTech dataset** [4]: includes both normal and abnormal events, covering 13 different scenarios in public spaces. It is designed for VAD and localization tasks, allowing researchers to develop models that can not only identify but also localize unusual events within the video frames.

3. **The Web Dataset** [54]: is a comprehensive collection of videos featuring various crowd scenes, such as pedestrian walking, marathon running, escape panics, protesters clashing, and crowd fighting. This dataset serves as a valuable resource for testing anomaly detection algorithms in crowd scenarios. Its diverse range of scenes enables researchers to evaluate and fine-tune their models for effective anomaly detection in real-world situations.

4. **GBA dataset** [121]: comprises videos captured at the Polytechnic School, of the University of Alcalá. This dataset focuses on a wide range of abnormal events, enabling researchers to train and test models capable of detecting various anomalies in diverse settings, in the wild. Moreover, as this dataset was made by our research gruop (initially also for HAR) could be improved in order to cover the more tricky and frequent anomalies that can appear in a normal indoor environment.

All datasets explained before are going to be used for training and/or testing the PhD proposals in the following chapter, and to show the proposals' results compared to those of other state of the art works.

3.2 UCF-Crime Dataset

The large-scale, comprehensive video dataset called UCF-Crime [5] was generated with the goal of enhancing research in the area of VAD, with a particular focus on identifying criminal activity in surveillance video footage. The dataset was created by University of Central Florida (UCF) academics and has gained popularity as a benchmark for assessing the efficiency and performance of VAD systems.

A sizable collection of 1,900 video clips, representing over 128 hours of footage, may be found in the UCF-Crime dataset. These videos were obtained from a variety of sources, such as security cameras, motion pictures, and YouTube videos. The dataset includes information on 13 distinct types of crimes, including Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. This wide variety of crime types makes the dataset more indicative of actual circumstances while also posing a considerable challenge to VAD techniques.

Figure 3.1 offers a comprehensive view of the UCF-Crime dataset, with sample frames from 14 different categories that represent a range of anomalies and normal events:

1. **'Abuse'** category presents frames of physical or aggressive interactions.
2. **'Arson'** includes potential instances of fires set intentionally.
3. **'Assault'** shows one person physically attacking another.
4. **'Burglary'** captures the act of breaking and entering or theft.
5. **'Explosion'** category contains frames from minor to large-scale destruction.
6. **'Fighting'** features physical confrontations between individuals.
7. **'Robbery'** depicts scenarios of theft, possibly under threat.
8. **'Shooting'** involves the use of firearms.
9. **'Stealing'** covers acts of theft such as pickpocketing.
10. **'Shoplifting'** is specifically about stealing from stores, different from broader theft.
11. **'Vandalism'** shows intentional property damage.
12. **'Arrest'** frames likely depict individuals being detained by authorities.
13. **'Road Accident'** includes traffic mishaps and vehicle collisions.
14. **'Normal'** category provides baseline frames of daily life without criminal activity.

This diverse dataset is crucial for developing computer vision systems that can accurately identify and differentiate between normal and irregular behaviors.

The UCF-Crime dataset has two categories for the videos: normal and aberrant. There are 1,050 video clips total in the standard videos, all of which show ordinary activities and situations devoid of any criminal or suspicious activity. On the other side, the aberrant videos include 850 video snippets that depict various illegal behaviours. Each atypical video clip that had one or more instances of illegal behaviour was personally annotated by the researchers. Researchers can assess the precision and efficiency of VAD algorithms in identifying and localising illegal activity inside the video frames thanks to these GT labels.

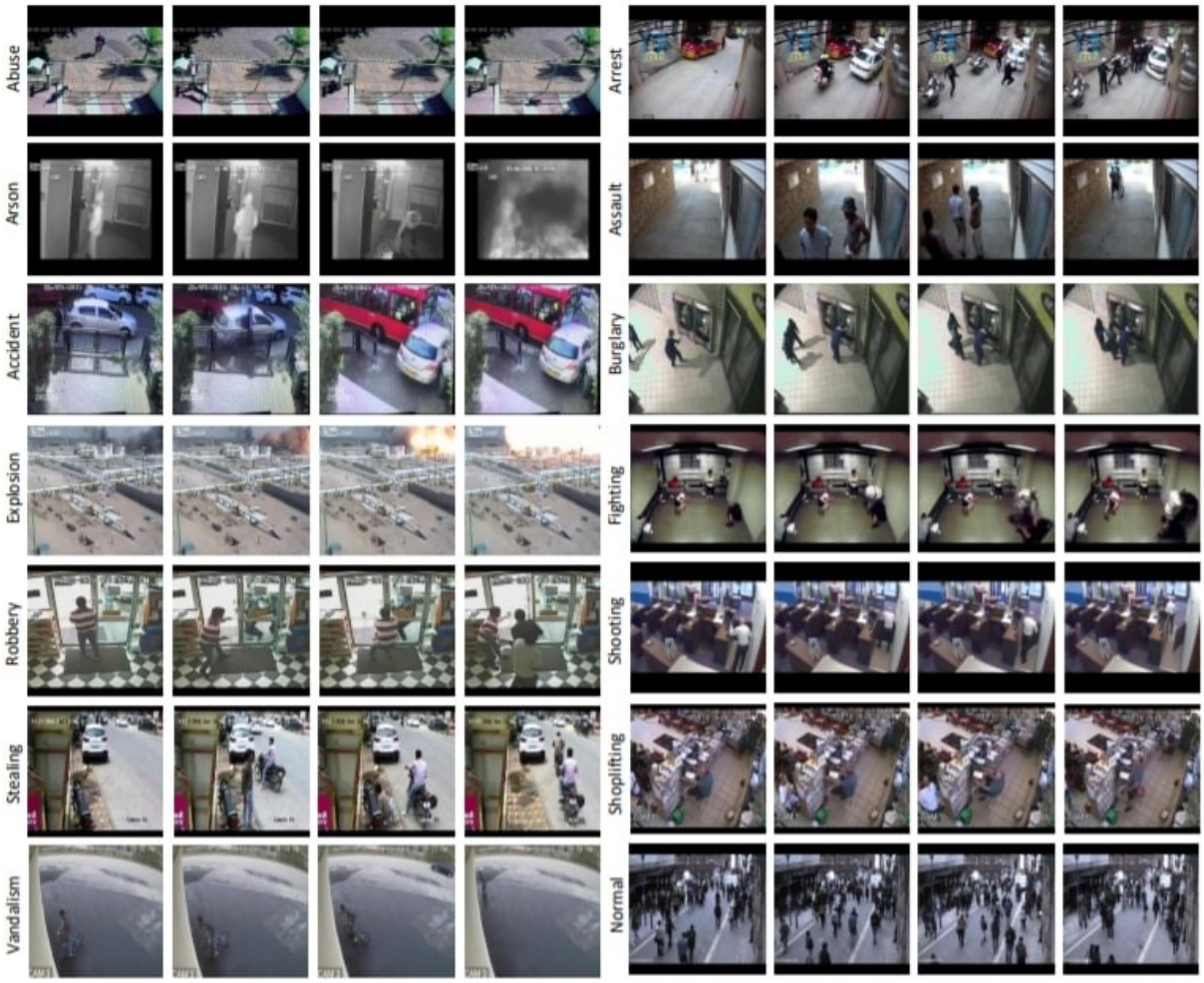


Figure 3.1: Sample frames from the different abnormal sequences selected from UCF-Crime dataset to be used in the PhD.

This dataset is distinguished by its complexity and diversity. The films in the collection have different lighting, camera angles, item sizes, and occlusion levels. This complexity aids in evaluating the robustness, generalizability, and situational awareness of VAD algorithms. Additionally, there is a lot of background motion in the dataset, including cars moving, people moving, and trees swaying in the wind. The dataset is a perfect benchmark for evaluating the limitations of VAD systems because these features make it more challenging to discriminate between normal and abnormal behaviour.

The UCF-Crime dataset's great level of annotation information is another noteworthy aspect. Each camera tape had aberrant events personally labelled by the researchers, who also provided accurate temporal information regarding the beginning and end times of the criminal actions. Researchers can use this data to evaluate the accuracy of VAD algorithms as well as their capacity to localise anomalies within the video timeline. Researchers can also conduct thorough performance analysis using the ground truth labels, including calculating precision, recall, and F1 scores as well as false positive and false negative rates.

3.3 ShanghaiTech Dataset

Another common benchmark for assessing VAD algorithms in the fields of computer vision and video analytics is the ShanghaiTech dataset [4]. Finding anomalous or unexpected patterns in data is the technique of VAD, which has applications in surveillance, intrusion detection, and industrial automation. Researchers have paid a lot of attention to the ShanghaiTech dataset, which focuses on VAD and features a variety of demanding settings and high resolution.

A team of researchers from ShanghaiTech University, lead by Dr. Yun-Fu Liu, produced the ShanghaiTech dataset. The dataset's main objective is to offer a thorough and accurate baseline for assessing the effectiveness of VAD algorithms as well as to promote the creation of new and improved methods. The dataset is downloadable from the university's official website and is accessible to the general public. 3.2 displays sample frames illustrating both ordinary and irregular sequences from the ShanghaiTech dataset.

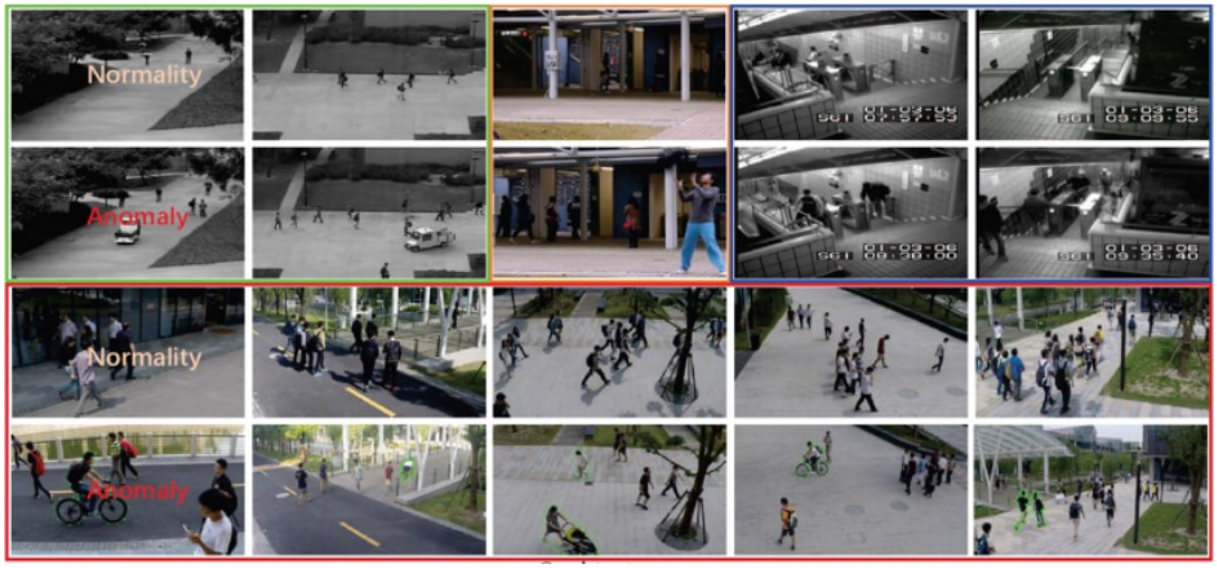


Figure 3.2: Sample frames from normal and abnormal sequences of ShanghaiTech dataset.

ShanghaiTech dataset includes two parts: Part A and Part B. A total of 437 high-definition video segments, spanning more than 13 hours, are present in both portions. The recordings were recorded using 13 distinct cameras that were placed throughout the ShanghaiTech University campus. Streets, parking lots, building entrances, and pedestrian areas are all covered by these cameras. The dataset is indicative of real-world scenarios thanks to the variety of scenarios, which also presents a considerable challenge to VAD systems.

The dataset's Part A consists of 330 video clips with a 1080p resolution and a 25 frames per second (fps) frame rate. Each segment in the videos has a runtime of about 1-3 minutes and was shot over the course of two months. There are 218 normal and 112 abnormal video clips in total in Part A. Anomalies of all kinds, including fights, loitering,

car theft, and unsecured baggage, can be seen in the abnormal video. The abnormalities are manually labeled by the researchers.

The dataset's Part B includes 107 video clips with a 720 pixels resolution and a 30 fps frame rate. Each clip in the videos has a runtime of about 1-3 minutes and was shot over the course of a month. There are 50 typical and 57 aberrant video clips in total in Part B. More subtle and difficult anomalies, such as strange walking gaits and people accessing forbidden locations, can be seen in the aberrant clips in Part B. Similar to Part A, the researchers manually documented and provided ground truth labels for the anomalies in this part.

As it has been previously mentioned, the ShanghaiTech dataset's high level of complexity, which presents a considerable challenge for VAD algorithms, is one of its distinguishing characteristics. The dataset contains a wide range of scenarios with various lighting, camera, and object size configurations. The videos also have a lot of background motion, like trees swinging in the breeze and individuals moving around the scene while walking or jogging, as well as several occlusions. These aspects make it challenging for algorithms to discern between typical and abnormal behaviour, and they aid in determining how robust and generalizable the techniques under consideration are.

The ShanghaiTech dataset's extensive annotation depth is another crucial feature. Each video clip's anomalous events were carefully marked by the researchers, who provided a precise temporal and spatial description of the abnormalities. This data can be used to assess the algorithms' accuracy in finding abnormalities as well as their capacity to localise those anomalies inside the video frame. Additionally, the ground truth labels allow researchers to conduct in-depth analysis of the algorithms' performance, including calculating precision and recall scores, false positive and false negative rates, and more.

3.4 The Web Dataset

The Web Dataset [54], introduced by Mehran *et al.*, is a benchmark dataset widely used for evaluating anomaly detection algorithms in crowd scenes. This dataset consists of 12 sequences of normal crowd scenes, including pedestrian walking, marathon running, and other everyday scenarios. Additionally, it contains 8 sequences of abnormal scenes, such as escape panics, protesters clashing, and crowd fighting.

The frames in The Web Dataset have been resized to a fixed width of 480 pixels for standardization. The dataset provides a comprehensive collection of both normal and abnormal crowd scenes, enabling researchers to develop and evaluate anomaly detection algorithms effectively.

Figure 3.3 displays sample frames from both the normal and abnormal sequences of The Web Dataset.

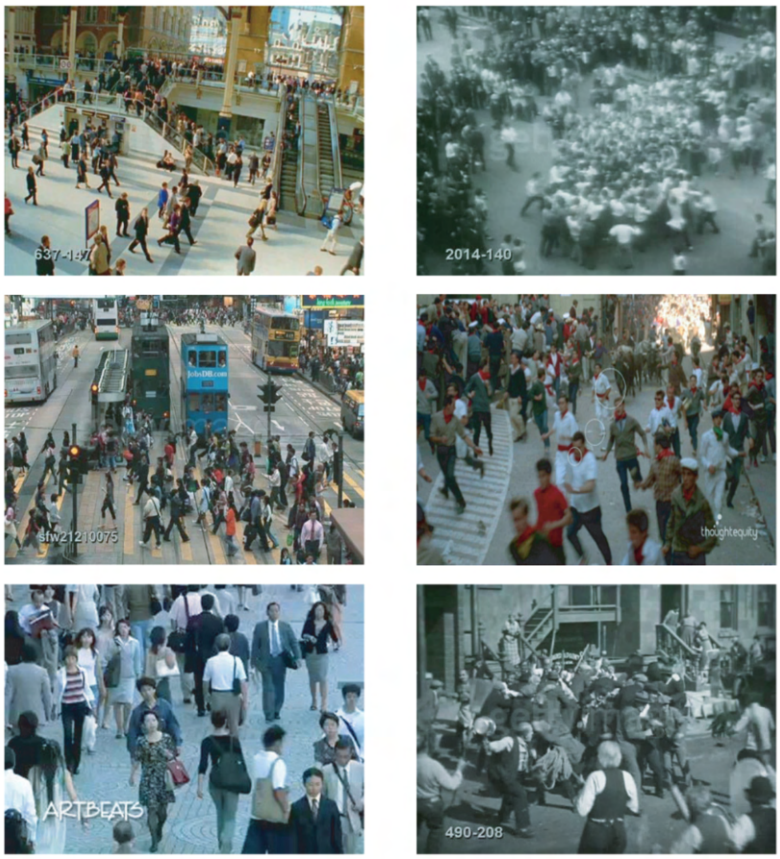


Figure 3.3: Sample frames from normal (left column) and abnormal (right column) sequences of The Web Dataset.

The inclusion of both normal and abnormal scenes in the dataset allows for a comprehensive evaluation of anomaly detection algorithms. Researchers can analyze the performance of their models in distinguishing between normal and abnormal crowd behaviors, providing valuable insights into the effectiveness of different anomaly detection techniques.

3.5 GBA Dataset

GBA dataset is made up of a series of videos shot by a fixed Full HD 2.7K camera, a GoPro HERO4, with a 1920 × 1080 pixel resolution, in a single, consistent environment. It has a broad field of view and records images at 50 fps, being the videos stored in `.MP4` format. The dataset presented here is a extension of GBA2016, initially described in [121]. The dataset is also available for access through the Geintra webpage [6].

3.5.1 Recording Setup

In order to capture the most area and minimise any occlusions, the camera was placed in the south of the Polytechnic School, as shown in figure 3.4. It was focused with a northwest orientation and in a zone with a particular height, roughly 1.8m.

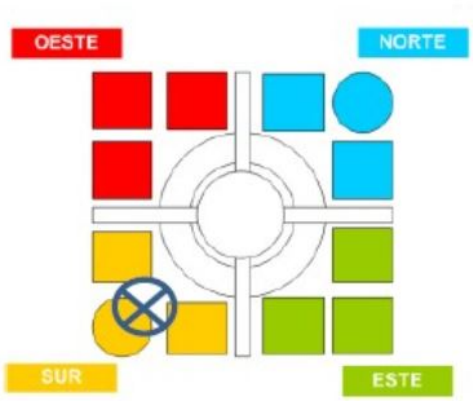


Figure 3.4: Top schematic view of the UAH's Polytechnics School.

The recording place is a hall next to some corridor with columns, stairs, an elevator, vending machines and a bench. The interesting points are shown in figure 3.5 that shows a map of the captured region, and in figure 3.6 that displays a view of the same are, highlighted in blue.

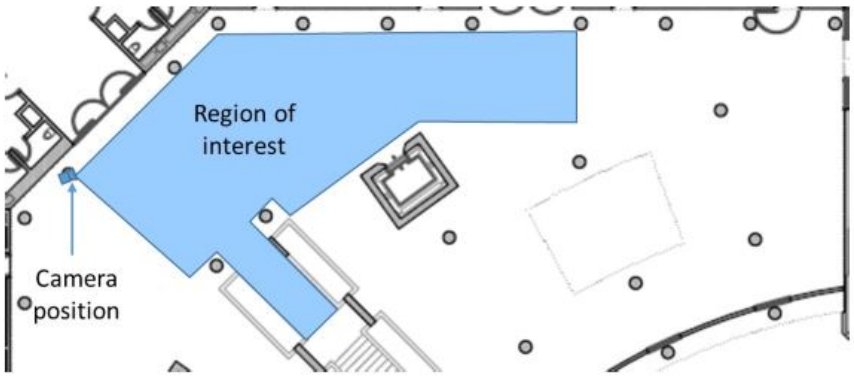


Figure 3.5: Region of interest in GBA dataset. Top view [121].

Due to the canteen being on the left side of figure 3.6, the library being above, and there being classroom hallways, the area of interest is also one of the busiest locations of the entire technological college. All of this opens up the prospect of numerous people showing up on the scene and engaging in various activities.

Additionally, because it is natural light, the illumination of the area of interest is dependent on the time of day, particularly, and it cannot be controlled. Due to this, the area of interest and consequently the films' scenario are uncontrolled, in the wild, and realistically and naturally occurring, which is essential for the goal of interior video surveillance.

3.5.2 GBA Characteristics

The GBA dataset, which was originally made up of 22 movies shot in 2016 [121], has been supplemented with 32 footage thanks to the participation of 17 individuals (also known

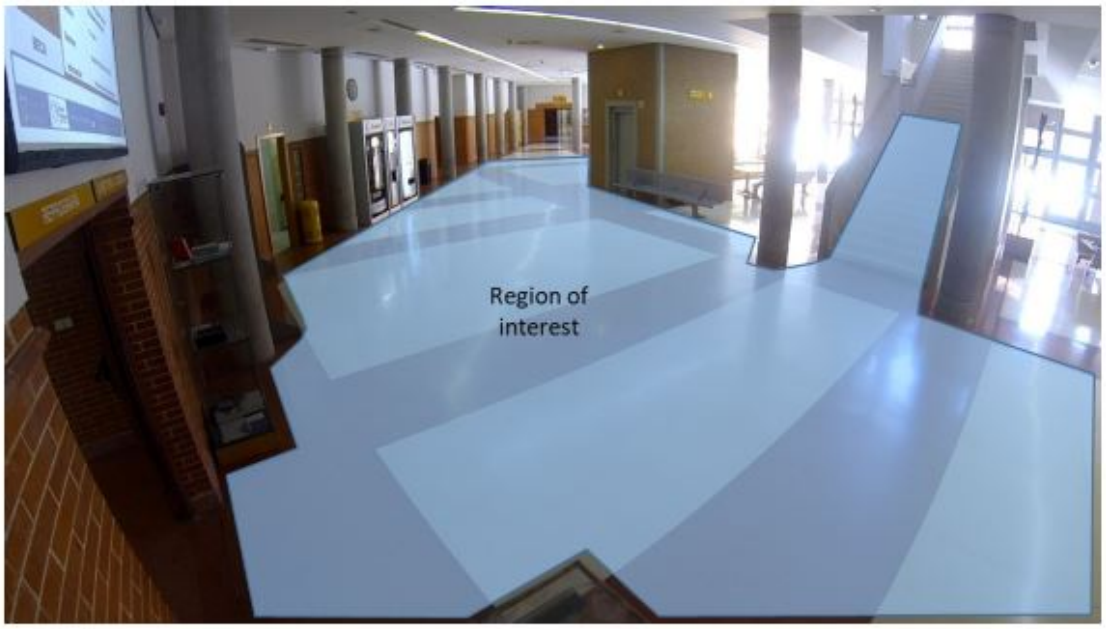


Figure 3.6: Region of interest in GBA dataset. Floor view [121].

as users), with three of them being women and the rest being men. There are 10 videos with unidentified intrusions in the scene. Each user is given an identification number in the dataset annotation, which is explained below, ranging from 1 to 20, whereas intruders are given a number between 1 and 100 to help identify them from identified users.

Each video's length ranges between 15min and 1min and 53s depending on the actions that occur in it.

In addition, 14 of the 17 users appear in two kinds of videos, with normal behaviours and anomalous ones:

- **Anomalous event ('Falling'):** in this first type, the person approaches the centre of the area of interest, in front of the stairs, and trips and falls. Then, the person gets up and walks back to where he was (see figure 3.7).

- **Normal event:** in the second type of videos, the subject first explores the area of interest before briefly disappearing through the door on the left side of the frame (as shown in figure 3.8). The person then starts walking till they reach the stairs, climbs them, descends them, and walks to the bench to sit down. After some time sitting down, the individual gets up and walks across the area till they reach the bench, where they take another seat (as in figure 3.9). The person then returns to the beginning place and begins running along the area of interest till reaching the lift. After that, they turn around and race back to the left side door until they reach the stairs (as shown in figure 3.10).

Figure 3.7: Sample images from an anomalous event in GBA dataset (**'Falling'**).

Figure 3.8: Sample images from a normal event (**'Walking'**) in GBA dataset.

Figure 3.9: Sample images from different normal events appearing in a single video (**'Walking'**, **'Sitting'** and **'Stairs'** actions) in GBA dataset.

Figure 3.10: Sample images from different normal events appearing in a single video (**'Running'** and **'Stairs'** actions) in GBA dataset.

Then, there is a last video (see figure 3.11) in which they appear 6 people, 2 women and 4 men, walking around the region of interest, sitting down in the bench and going upstairs and downstairs.

Figure 3.11: Sample images from a normal event in GBA dataset with multiple people.

The actions included in the original dataset are five: **'Walking'**, **'Running'**, sitting down (called **'Sitting'**), falling down (called **'Falling'**) and climbing stairs (named **'Stairs'**).

Moreover, the completed dataset includes 12 new videos including groups of people (between ten and thirty individuals) recreating another anomalous situation specifically recorded for this PhD. People in such videos emulate feeling an alarming situation and start suddenly to run in the same and different directions to escape from that. The situation is very similar to that in UMN dataset, we called it **'Stampede'**, and it was recorded in the same area and with the same camera that the original GBA.

Table 3.1 describes the content and distribution among HAR sequences in the GBA dataset finally created in this PhD, and made public for the scientific community in [6].

Table 3.1: GBA dataset overview detailing the number of individuals and sequences per action.

Action	**Individuals**	**Sequences**
Walking	13	72
Running	12	75
Sitting	16	78
Falling	14	70
Stampede	10-13 simultaneously	12

In order to train and test the action recognition system proposed, it is necessary to select several sequences to be used to train it and then use the rest to test the training. Thus, each video is split into pieces, to be used in the training and the test of the HAR developed.

3.6 Summary of Datasets

In this chapter, we explored a variety of datasets essential in the field of anomaly detection, each significant for HAR and VAD within the realms of ML and DL.

UCF-Crime dataset is notable for its extensive collection of 1,900 video clips, encompassing over 128 hours of footage. This dataset includes a diverse range of resolutions, sourced from security cameras, motion pictures, and online platforms. It covers 14 different types of crime categories, making it a comprehensive resource for studying criminal activities and normal scenarios.

ShanghaiTech dataset contributes with 437 high-definition video segments, totaling more than 13 hours. Captured at 1080 pixels and 720 pixels of resolution, it covers various scenarios in public spaces and is designed for VAD tasks, featuring a mix of normal and abnormal events.

The Web Dataset offers a unique focus on crowd behaviors with 20 video sequences, including both normal and abnormal crowd scenes. The videos have been standardized to a resolution of 480 pixels. This dataset is instrumental for analyzing a range of crowd behaviors, from pedestrian activities to more dynamic scenes such as protests or crowd panics.

Finally, **GBA dataset** consists of 54 videos of varied lengths, recorded in Full HD (1920×1080) resolution. Focused on indoor environments, with a limited number of individuals, it captures a variety of indoor activities, encompassing both normal and anomalous events, and provides a specialized resource for indoor surveillance studies.

These datasets collectively offer a diverse spectrum of scenarios, resolutions, and event types, crucial for training and evaluating the efficacy of HAR and VAD systems. The subsequent chapters will utilize these datasets to demonstrate the proposed methodologies and compare their performance against existing state of the art works.

Chapter 4

First contribution: Using Weakly Labeled Training Videos for Detecting Abnormal Human Behaviour in Video-Surveillance Scenes

Article

Semi-Supervised Anomaly Detection in Video-Surveillance Scenes in the Wild

Mohammad Ibrahim Sarker, Cristina Losada-Gutiérrez *, Marta Marrón-Romera, David Fuentes-Jiménez and Sara Luengo-Sánchez

Department of Electronics, Politechnics School, Campus Universitario S/N, University of Alcalá, Alcalá de Henares, 28801 Madrid, Spain; ibrahim.sarker@uah.es (M.I.S.); marta.marron@uah.es (M.M.-R.); d.fuentes@edu.uah.es (D.F.-J.); sara.luengo@edu.uah.es (S.L.-S.)
* Correspondence: cristina.losada@uah.es; Tel.: +34-91-885-6906

Abstract: Surveillance cameras are being installed in many primary daily living places to maintain public safety. In this video-surveillance context, anomalies occur only for a very short time, and very occasionally. Hence, manual monitoring of such anomalies may be exhaustive and monotonous, resulting in a decrease in reliability and speed in emergency situations due to monitor tiredness. Within this framework, the importance of automatic detection of anomalies is clear, and, therefore, an important amount of research works have been made lately in this topic. According to these earlier studies, supervised approaches perform better than unsupervised ones. However, supervised approaches demand manual annotation, making dependent the system reliability of the different situations used in the training (something difficult to set in anomaly context). In this work, it is proposed an approach for anomaly detection in video-surveillance scenes based on a weakly supervised learning algorithm. Spatio-temporal features are extracted from each surveillance video using a temporal convolutional 3D neural network (T-C3D). Then, a novel ranking loss function increases the distance between the classification scores of anomalous and normal videos, reducing the number of false negatives. The proposal has been evaluated and compared against state-of-art approaches, obtaining competitive performance without fine-tuning, which also validates its generalization capability. In this paper, the proposal design and reliability is presented and analyzed, as well as the aforementioned quantitative and qualitative evaluation in-the-wild scenarios, demonstrating its high sensitivity in anomaly detection in all of them.

Keywords: anomaly detection; RGB; CNN; multiple instance learning; video-surveillance

check for updates

Citation: Sarker M.I.; Losada-Gutiérrez, C.; Marrón-Romera, M.; Fuentes-Jiménez, D.; Luengo-Sánchez, S. Semi-Supervised Anomaly Detection in Video-Surveillance Scenes in the Wild. *Sensors* **2021**, *21*, 3993. https://doi.org/10.3390/s21123993

Academic Editor: Paolo Spagnolo

Received: 30 April 2021

In this chapter of the PhD thesis, we delve into the critical use of surveillance cameras to enhance public safety. The aim is not only to present a summary of the research conducted but also to provide a comprehensive introduction to the topic, shedding light on the issues associated with manual monitoring and the inherent challenges posed by human error due to exhaustion.

In modern society, using surveillance cameras has become more and more common, greatly enhancing public safety and reducing crime. However, the success of these systems

frequently depends on their capacity to accurately monitor massive volumes of video data. Although necessary, manual monitoring can be exhausting and prone to mistakes as human operators grow weary, perhaps resulting in gaps in surveillance coverage.

This chapter addresses these concerns by introducing an innovative approach to autonomous anomaly detection in surveillance videos. The primary focus of this work is the utilization of a weakly supervised learning algorithm known as MIL. With this algorithm, manual monitoring can be replaced with the automatic detection of anomalies in surveillance video, which is a more effective and trustworthy method.

The research uses a pre-trained temporal 3D CNN and spatio-temporal data taken from each video source to do this. This cutting-edge technology makes it possible to extract complex patterns and subtleties from the video data, making it easier to classify video clips as 'normal' or 'abnormal'. We want to lower the possibility of human error and guarantee a constant level of surveillance coverage by automating this process.

In addition to introducing the MIL algorithm and employing neural networks, this chapter also delves into an Enhanced ranking loss function. This refined approach plays a crucial role in optimizing the efficacy of the system by effectively enhancing the differentiation between normal and abnormal videos. By implementing this strategy, it assists in reducing the occurrence of false negatives, a critical factor in ensuring the reliability of a surveillance system.

As we progress further into this chapter, we will not only examine the technical components of the suggested methodology but also explore the practical ramifications and prospective applications of this autonomous anomaly detection system.

4.1 Introduction

The primary objective of this study is to enhance public safety by addressing a range of potential threats, such as robberies, accidents, and antisocial behavior. The specific focus of this research is the automation of anomaly detection in surveillance sequences. In this context, conventional monitoring methods, predominantly reliant on human processes, necessitate substantial effort, are laborious, and are susceptible to errors. There is a need to transition towards employing automated anomaly detection methods that utilize computer vision techniques in order to promptly identify and categorize abnormal occurrences.

As explained in chapter 2, previous efforts in automatic anomaly identification have primarily focused on specific anomaly detectors or employed unsupervised or supervised techniques, both of which possess inherent limitations. Supervised methods necessitated substantial annotations for training purposes, but unsupervised approaches were susceptible to false positives due to the variability of normal events.

This chapter presents an improved method for anomaly detection based on MIL, a weakly supervised learning methodology, to provide increased performance and superior generalization. The proposed method contributed was initially described in [23]. The following are the key contributions of this study:

1. Use of a T-C3D [1] trained on a sizable, noise-free Kinetics dataset [122] for feature extraction.
2. Creation of a modified ranking loss function that substantially decreases the miss rate, increasing anomaly detection performance.
3. A thorough proposal based on MIL is presented, displaying great generalization potential for finding anomalies in hypothetical real-world situations.

The proposed architecture for anomaly detection in surveillance videos using MIL is presented in figure 4.1.

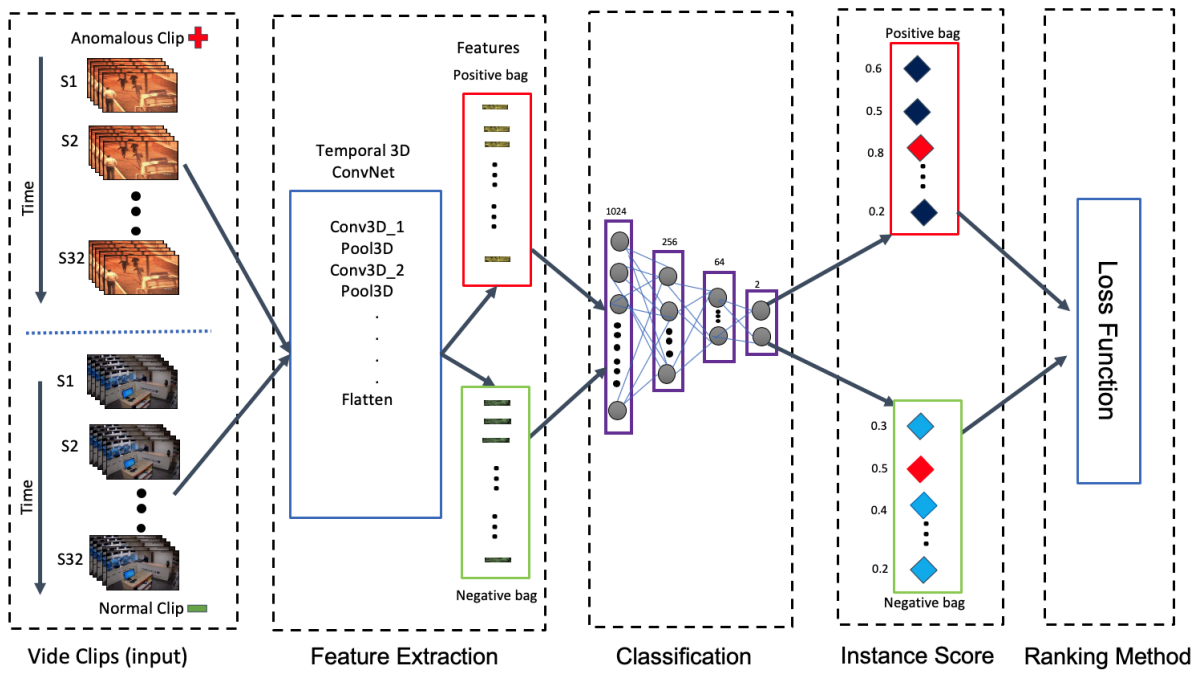


Figure 4.1: Proposed architecture for anomaly detection in surveillance videos.

Each of the stages shown in figure 4.1 are explained in detail in the following sections.

4.2 Methodology

The proposed methodology for anomaly detection in surveillance videos using MIL is discussed exhaustively in this section. Our proposed architecture, depicted in figure 4.1, serves as a response to the increasing demand for efficient and reliable anomaly detection systems in surveillance contexts.

4.2.1 Proposed Architecture and Training Process

4.2.1.1 Input data pre-processing

A key initial step in the weakly supervised anomaly detection framework for video surveillance involves processing each video sequence from a comprehensive dataset, the UCF-Crime [5]. This dataset, which is full with examples of criminal activity in the actual world, offers a stable learning environment for our model.

Deep learning models frequently encounter challenges while attempting to process lengthy video sequences in their entirety, primarily due to their high computational complexity. This assertion has special validity when dealing with high-definition surveillance footage. Therefore, during the processing stage, each video input from the UCF-Crime dataset is divided into 32 temporal segments, that do not overlap. The aforementioned segments are commonly acknowledged as weakly supervised bags and are employed as training instances for the model.

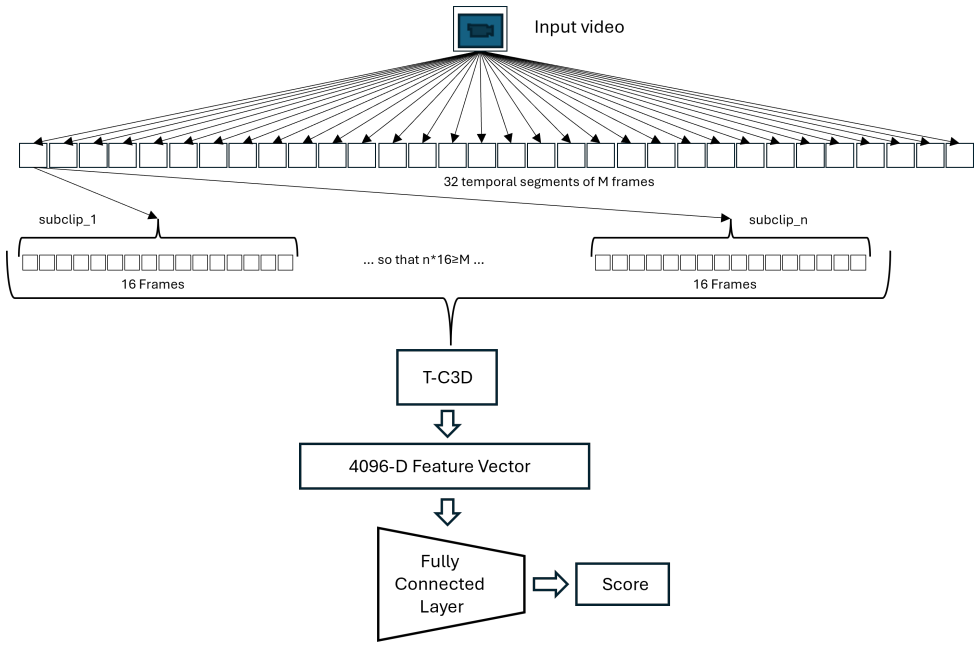


Figure 4.2: Workflow of video segmentation.

The input video is divided into 32 temporal segments, each containing M frames. Within each of these segments, the model further breaks down the content into multiple subclips, each consisting of 16 frames. This structural organization is crucial for capturing the temporal dynamics necessary to effectively identify anomalous behavior (see figure 4.2). Each segment is composed of several subclips (denoted as n subclips), which are processed through a pre-trained 3D convolutional neural network (CNN), specifically the T-C3D model. This model excels at interpreting both spatial arrangements and temporal movements within the video data. It is ensured that $n \times 16 \geq M$ meaning the total frames across all subclips are equal to or exceed the total frames in the segment allowing comprehensive coverage and analysis.

Feature extraction is performed on each subclip, and the extracted feature vectors from all subclips within a segment are averaged to produce a single 4096 dimensional feature vector. This vector effectively encapsulates the collective visual and temporal information of the entire segment, providing a comprehensive representation for anomaly detection. The 4096 dimensional feature vector for each segment is then fed into a fully connected layer which outputs a score.

Thus, The choice was reached after conducting numerous experiments, benchmarking, and considering the current best practices documented in the modern literature. It is important to highlight that our adaptive segmentation approach has the potential to be subdivided into 64 segments for longer videos, hence increasing the level of detail in feature extraction.

In addition, the temporal segmentation technique was carefully designed to correspond with the distinctive time frames of the detected abnormalities or criminal actions that are the focus of our study. The careful selection of data in our model prevents it from falling into the trap of temporal myopia, where it ignores short-lived anomalies, as well as the trap of temporal hyperopia, where it excessively breaks down long-lasting irregularities, thus reducing their importance.

The segmentation process employed functions as a crucial temporal filter, effectively synchronizing with the inherent temporal patterns of the underlying data. The alignment in question allows the model to effectively identify and analyze detailed spatio-temporal patterns inside each segment. This capability enables the model to accurately detect deviations occurring at different temporal scales. The intentional synchronization described here is crucial, as it ensures that the model remains attentive to the nuances of temporary deviations while preserving the contextual consistency of longer-lasting aberrations.

In summary, our segmentation methodology is a crucial aspect of our model's ability to effectively identify abnormal occurrences in video surveillance data. This serves as evidence of our dedication to utilizing the inherent temporal organization of the data, guaranteeing that our DL framework excels in detecting variations over a range of temporal complexities, hence enhancing its effectiveness and resilience in practical applications in the real world.

The model has the capability to enhance its ability to detect abnormal or criminal activities by segmenting video sequences into 32 segments, hence focusing on localized information. While the effectiveness of using 32 segments for our specific assignment has been observed, it is important to acknowledge that the segmentation approach may require modification depending on the specific requirements of the activity at hand.

The resultant model capitalizes on limited supervision by effectively processing large, unlabeled datasets and effectively addressing the intricacies of video data by incorporating temporal patterns at an appropriate scale. Hence, this approach offers a dependable and

efficient means of detecting anomalies or illicit behavior inside video surveillance data, showcasing significant prospects for practical use.

After segmenting the video, the subsequent phase in the weakly supervised anomaly detection framework involves generating positive and negative bags from these segments. The utilization of the distinctive elements of our methodology in this step is of utmost importance for the model's ability to differentiate between typical and aberrant behaviors in video data.

The negative samples consist of segments extracted from conventional or representative crime video within the collection. The occurrences in question serve as a benchmark or reference point for identifying abnormalities, representing behavior or activity that is considered typical. In contrast, positive bags consist of components derived from atypical or deviant visual recordings, specifically movies that capture occurrences of illicit behavior. These excerpts function as examples of non-traditional conduct.

In order to facilitate model training, it is imperative to meticulously construct these bags. The model is subjected to a range of typical and atypical patterns, so producing a spectrum that encompasses both conventional and deviant behaviors. The present study introduces a contrastive learning framework. The fundamental basis of anomaly detection lies in the capacity to identify deviations from standard behavior, a capability facilitated by the utilization of this framework.

It is important to acknowledge that although there is some degree of loose control over these bags, not every occurrence inside the positive bag can be considered remarkable, and conversely, not every incidence within the negative bag can be deemed entirely regular. This characteristic poses challenges to the learning process as the model needs to acquire the ability to differentiate between typical and atypical behavior, while also determining the level of uncertainty associated with each instance.

Through the development of this framework, we want to provide a comprehensive understanding of both conventional and atypical behaviors within the given paradigm. The model enhances its understanding of normal and abnormal patterns by acquiring the ability to evaluate video sequences and their segments in relation to these categories. Consequently, this capability enables the system to effectively detect unlawful behavior or irregularities in newly captured video footage, thereby significantly augmenting its practical significance and usefulness.

4.2.1.2 Feature extraction

One of the cornerstones of our methodology is feature extraction, a task entrusted to the T-C3D architecture [1]. The T-C3D employs a hierarchical method in order to effectively understand the intricate activities depicted in video. The proposed approach effectively captures the temporal dynamics of the video while simultaneously preserving the spatial

context by treating the spatial and temporal components as distinct yet interconnected entities inside a unified model. In contrast to alternative designs that operate on a per-frame basis for video processing, the T-C3D model adopts a batch-wise approach, utilizing a deliberate aggregation function to extract features at the video level. The implementation of this approach not only guarantees a notable degree of precision but also significantly enhances the capacity for real-time inference.

Moreover, in our experiments, we compared T-C3D with another popular feature extractor, C3D, as used in [5]. We found that T-C3D outperforms C3D in terms of performance metrics such as accuracy and computational efficiency. The motivation behind choosing T-C3D over C3D lies in its superior ability to capture both temporal and spatial features simultaneously, leading to more robust representations of video data for our specific task.

Pre-training neural networks is a widespread tactic to imbue them with a preliminary understanding of the data domain. Studies leveraging Long-term Temporal Convolution (LTC) [123] have previously pre-trained on the Sports-1M dataset [124] for this. While rich in volume, Sports-1M suffers from noise due to its auto-generated annotations. On the other hand, the Kinetics dataset, introduced by [122], is both vast and meticulously annotated, covering a diverse range of human actions across categories. Pre-training the T-C3D with the Kinetics dataset instills it with a refined comprehension of video data, subsequently leading to improved anomaly detection.

4.2.1.3 Classification

After the process of feature extraction, the subsequent task that arises is classification. The feature vector produced by T-C3D is subjected to processing by our own classifier network. The classification of returned feature vectors as positive (indicating the presence of an anomaly) or negative (indicating the absence of an abnormality) is achieved by the utilization of several components in the system.

Thus, the classifier network, also known as the deep MIL network, is located directly following the T-C3D. It is a complex structure consisting of multiple layers, including a 3D average pooling layer and four fully connected layers. Each of these layers is enhanced by batch normalization. The deliberate incorporation of a dropout mechanism at a rate of 60% into the Rectified Linear Unit (RELU) activation function is a strategic decision aimed at improving the generalization of a model by mitigating the risk of overfitting.

The nature of MIL makes it particularly suitable for situations like ours, when obtaining detailed annotations is either not possible or not feasible. The process of training with MIL entails the creation of bags comprising instances, specifically video segments in this context. The implementation of a high-level approach to labeling significantly mitigates the need for extensive human effort and minimizes the likelihood of errors that are inherent in the process of frame-by-frame annotations. The training method is constructed based

on MIL principles. It subsequently constructs models for these collections of instances, acquiring knowledge about the complexities involved in identifying anomalies.

Conventional binary classification methods encounter challenges in meeting the level of specificity required for anomaly identification. Therefore, we redefine the problem as a regression task, with emphasis on the disparity in score between regular and abnormal segments . The model has been trained to prioritize anomalous segments by assigning them significantly higher scores compared to their typical counterparts, as seen by the following relationship:

$$f(\mathcal{V}_a^i) > f(\mathcal{V}_n^i), \tag{4.1}$$

where $f(\mathcal{V}_a^i)$ and $f(\mathcal{V}_n^i)$ represents the normalized scores for anomalous and normal i video segment, respectively.

The normalization of scores by f is achieved through the application of a sigmoid function in the final layer of the model, transforming the raw output scores to a range between 0 and 1. This transformation is crucial for comparing these scores against a threshold to classify segments as anomalous or normal, efficiently using video-level labels for model training. In this way, scores from $f(\mathcal{V}_a^i)$ closer to 1 indicate a higher likelihood of an anomaly, whereas $f(\mathcal{V}_n^i)$ scores closer to 0 signify normality.

Labels, i.e. GT, apply to the entire video rather than specific segments within it. This indicates that while a video may be categorized as containing an anomaly, its precise temporal location within the video remains unspecified. The use of normalized scoring functions and ranking losses in this MIL setting enhances learning from such weakly labeled data. It optimizes the model's ability to detect anomalies by focusing on maximizing the separation between the highest anomaly scores in positive (anomalous) bags and those in negative (normal) ones.

It is important to remember that the labels, also known as ground truths, apply to the entire video rather than specific segments within it. This indicates that while a video may be categorized as containing an anomaly, the precise location or occurrence of the anomaly within the video remains unspecified. The employment of normalized scoring functions and ranking loss in this MIL setting aids in learning from weakly labeled data. It optimizes the model's ability to detect anomalies by focusing on maximizing the separation between the highest anomaly scores in positive (anomalous) bags and the scores in negative (normal) bags.

4.2.1.4 Model training

Model optimization is pivotal in DL. AdaGrad [125] remains a popular choice, it tended to aggressively reduce learning rates in our context. Adadelta [126], on the other hand,

showcased more stability. The Adam algorithm [127], though robust in many applications, was observed to have momentum minimization issues in our specific use case.

4.2.2 Proposed Ranking Loss Function

Anomaly detection, especially in the context of videos, requires a nuanced approach to identifying abnormal segments. The importance of loss functions, particularly in this domain, is pivotal. Among the proposed methodologies, the one presented by Sultani et al. [5] has been seminal.

Sultani suggested a distinctive method to aggregate classification scores within a bag. This method is encapsulated by the MIL ranking loss, succinctly represented in equation 4.2.

$$\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) > \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right) \tag{4.2}$$

This equation elucidates a simple yet profound concept. Each bag is attributed with the highest score derived from its constituent video segments. This underpins the essence of the ranking objective: segments with the most pronounced anomalies (visualized in red, as exemplified in figure 4.1) should inherently supersede those from the non-anomalous or negative bags (depicted in green).

Building on this foundational idea, the proposal in Sultani *et al.* [5] went on to recommend employing the ranking loss function expressed in equation 4.3 during the decisive classification phase.

$$\mathcal{L}\left(\mathcal{B}_a, \mathcal{B}_n\right) = \max\left(0, 1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) \tag{4.3}$$

While the existing methodology offers a robust foundation, this work seeks to introduce refinements to enhance precision. By incorporating the $\log_2$ function, our proposed ranking loss function targets minimization of variability in anomalous sequences, thereby accentuating the distinction between scores associated with normal and abnormal sequences, as expressed in equation 4.4.

$$\begin{aligned}\mathcal{L}\left(\mathcal{B}_a, \mathcal{B}_n\right) = &\max\left(0, 1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) \\ &+ \max\left(0, \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right) - \log_2\left(\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right)\right)\right)\end{aligned} \tag{4.4}$$

where:

- $\mathcal{L}\left(\mathcal{B}_a, \mathcal{B}_n\right)$ represents the proposed ranking loss function, where $\mathcal{B}_a$ and $\mathcal{B}_n$ are the anomalous and normal bags respectively.

- The term $\max\left(0, 1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right)$ is a hinge loss component. Within this, $1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)$, signifies a margin-based loss where:
  - $\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right)$: Represents the highest anomaly score from the anomalous bag $\mathcal{B}_a$.
  - $\max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)$: Represents the highest anomaly score from the normal bag $\mathcal{B}_n$.
- The component $\max\left(0, \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right) - \log_2\left(\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right)\right)\right)$ introduces the $\log_2$ function. This term minimizes the variability in anomalous sequences and accentuates the distinction between scores of normal and abnormal sequences. By subtracting the log-transformed highest anomaly score from the anomalous bag from the highest anomaly score of the normal bag, it provides a kind of normalization and seeks to amplify differences between normal and anomalous sequences.

The primary goal of this enhanced function is to systematically reduce the miss rate. Ensuring that anomalous cases are not mistakenly classed as normal is of paramount importance in anomaly detection applications. Given the critical nature of many applications in this domain, such misclassifications could lead to significant setbacks and costs.

Previous approaches often overlooked the temporal nuances intrinsic to anomalies. Specifically, anomalies, by their inherent nature, frequently appear within short temporal spans, leading to a lack of temporal structure in the recordings labeled as anomalous. An effective anomaly score should encapsulate the continuous narrative of video data, providing temporal coherence.

To address this, the introduced function integrates a regularization component to account for variations in anomaly scores between consecutive video frames, consistent with the methodologies outlined by [128].

Considering both temporal coherence and feature sparsity comprehensively, the improved ranking function presented in this research is formulated as in equation 4.5:

$$\begin{aligned} \mathcal{L}\left(\mathcal{B}_a, \mathcal{B}_n\right) = & \max\left(0, 1 - \max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) \\ & + \max\left(0, \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right) - \log_2\left(\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right)\right)\right) \\ & + \lambda_1 \sum_{i}^{n-1} \left(f\left(\mathcal{V}_a^i\right) - f\left(\mathcal{V}_a^{i+1}\right)\right)^2 + \lambda_2 \sum_{i}^{n} f\left(\mathcal{V}_a^i\right) \end{aligned} \tag{4.5}$$

wherein:

- λ_1 and λ_2 are regularization coefficients, fine-tuned typically via cross-validation. Specifically: λ_1 accentuates the importance of temporal continuity. A heightened value reinforces the preservation of temporal coherence between consecutive video

segments. While λ_2 emphasizes feature sparsity. A magnified value prompts the model to possess a sparser anomaly score representation.

- $f\left(\mathcal{V}_a^i\right)$ quantifies the anomaly score of segment $\mathcal{V}_a^i$.
- The term $\sum_i^{n-1}\left(f\left(\mathcal{V}_a^i\right)-f\left(\mathcal{V}_a^{i+1}\right)\right)^2$ enforces temporal smoothness by penalizing abrupt score variations between successive segments.
- $\sum_i^n f\left(\mathcal{V}_a^i\right)$, when weighted by λ_2, propels the model towards a sparse anomaly score landscape.

This advanced loss function not only distinguishing between regular and irregular segments within a video but also acts as a temporal analyzer, adding a critical time-based dimension to the analysis. It doesn't just flag anomalies, it timestamps them. Furthermore, it offers a streamlined, focused, and actionable representation of the video data, providing clarity and simplifying the task of identifying and addressing anomalies. This is a significant advancement in the field of DL, particularly in applications where timely anomaly detection is of utmost importance, such as surveillance, security, and quality control.

4.3 Experimental Results and Discussion

4.3.1 Experimental Setup

To ensure a comprehensive evaluation, the method was tested on three distinct datasets: UCF-Crime [5], GBA [121], and The Web Dataset [54]. The UCF-Crime dataset serves a dual purpose, being utilized for both training and testing the proposed system. However, the GBA and Web Dataset were exclusively utilized for testing purposes. The purpose of strategically implementing this strategy was to evaluate the efficacy of the solution in different situations and when faced with unexpected data samples. More detailed insights into the attributes of these databases can be found in chapter 3.

The performance of the proposed model is evaluated by utilizing the Receiver Operating Characteristic Receiver Operating Characteristic (ROC) curve. This graph illustrates the relationship between the true positive rate and the false positive rate at various thresholds. It allows for a precise evaluation of a model's capacity to differentiate between normal and abnormal occurrences. The curve's adaptability stems from its ability to provide insights into the appropriate selection of thresholds, taking into account the desired balance between detecting anomalies and minimizing false alarms. Furthermore, the area under the curve Area Under the Curve (AUC) offers a measurable indicator for directly comparing models, making the receiver operating characteristic ROC curve a crucial tool for evaluating and assuring strong performance in various surveillance scenarios. Table 4.1 provides a performance comparison of the proposed model with that of

the models proposed in [5], [114], [109] and [116], displaying as classification quality value the AUC for each proposal tested in UCF-Crime dataset.

4.3.2 Performance Evaluation

The experiments for anomaly detection were rigorously carried out utilizing the comprehensive set of test videos present in the UCF-Crime dataset. In this context, the proposed methodology was able to achieve an AUC of 80.36%.

Interestingly, a variant of the proposed model, which operates without the inclusion of the $\log_2$ transformation, managed to secure an AUC of 78.63%. This score is almost analogous to the best performances exhibited by other existing state of the art methods, which cap at an AUC of 78.66%. Such results not only vouch for the effectiveness of the proposed methodology but also highlight its superiority and potential as a reliable tool for video-based anomaly detection. It can also be observed in figure 4.3, which displays the ROC comparison of the proposed approach (in green) as well as that of earlier methods.

In order to evaluate the generalization capability of the proposal, figure 4.4 shows the ROC curve obtained with The GBA and Web datasets without fine-tuning. In our experiment with the Web and GBA dataset, we resized all the test videos into 320*240 pixels (.mp4, H.264 codec) and manually annotated abnormal videos. The ROC in figure 4.4 shows that our proposal for The Web Dataset achieved an AUC of 78.58% which outperformed state of the art which was an AUC of 73%, as stated in the dataset publication [54], and for GBA dataset we have obtained 61.65%.

Table 4.1: Performance comparison using UCF-Crime dataset. The performance of the proposal appears in italics.

Model	**AUC (%)**
Binary classifier	49.99
Hassan *et al.* [114]	50.66
Lu *et al.* [109]	65.51
Sultani *et al.* [5]	75.41
Zhang *et al.* [116]	78.66
Proposed method without $\log_2$	***78.63***
Proposed method with $\log_2$	***80.36***

It is noteworthy that, our results on The Web Dataset are better than in GBA because some of abnormal classes in that dataset are similar to those of UCF-Crime dataset. On the other hand GBA dataset has no crime scene, which makes the model less predictable.

4.3.3 Qualitative Results

Having reported the quantitative results, qualitative ones are presented in figures 4.5, 4.6, 4.7, and 4.8. In order to visualize the improved performance of the the proposed method,

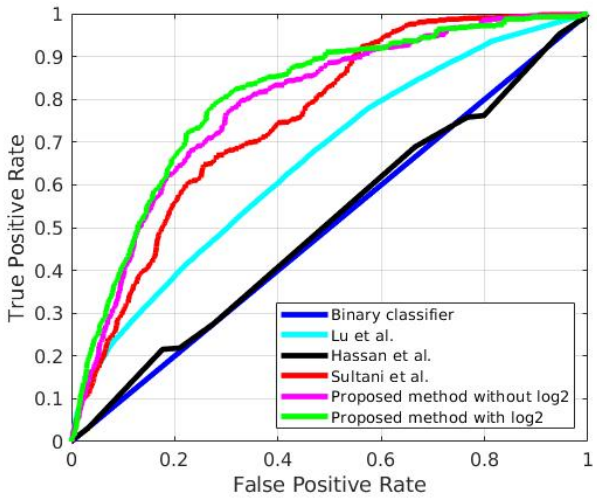


Figure 4.3: ROC comparison of a binary classifier (blue), Lu *et al.* [109] (cyan), Hassan *et al.* [114] (black), Sultani et al. [5] (red) and the proposed method (pink and green) in UCF-Crime dataset.

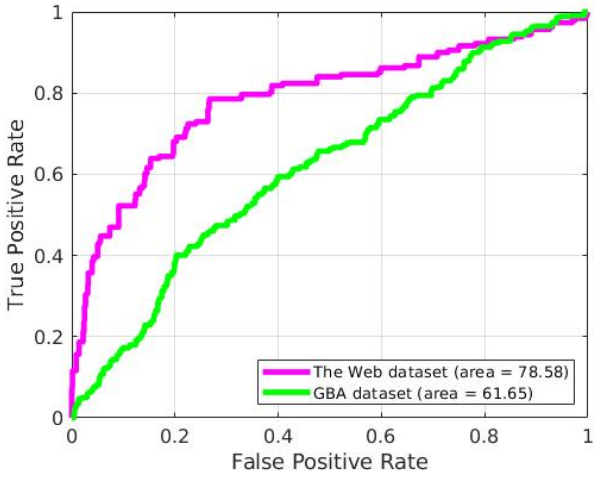


Figure 4.4: ROC of the proposed method in GBA and The Web Dataset.

both UCF-Crime dataset and real scenarios unseen by the algorithm from GBA dataset are considered. The scenarios comprise both far-view and near-view video-sequences.

The proposed methodology as well as the model proposed by Sultani *et al.* in [5] are executed on all of them, and the anomaly responses from both the models are compared. The anomaly ground truth of each video is then plotted along with the anomaly score obtained from each proposal. The mentioned anomaly ground truth indicates a maximum anomaly score with 1, while represents normal with a minimum of 0.

For each result, they are shown, in the corresponding figure, four images from each selected sequence at the top, and the anomaly scores given by the two different architectures at the bottom, within the ground truth (in dotted green): Sultani's one at the left side of the figure, and our proposal at the right side.

The first test (shown in figure 4.5) is performed over GBA dataset, corresponding to a far-view sequence in which a person falls down. It comprises four scenarios namely: (a) a person walking, (b) the person falling down, (c) the person still lying on the ground and (d) the person getting up. According to the ground truth, scenarios (b), (c) and (d) are considered anomalous while scenario (d) is normal. Within this context, as it can be seen in figure 4.5, the proposed methodology yields high anomaly scores for the scenarios presenting the person's fall ((b) and (c)) and getting up (d), whereas a low anomaly score for the normal event (person walking) in scenario (a). Besides, the model from Sultani gives a moderate score for scenario (b) and a low anomaly score for all the remaining three scenarios (a,c,d), what supposes the appearance of false negative. Thus, comparing the anomaly scores, the values assigned to scenario (b) by the proposed model are significantly high (near the maximum), so that it can be concluded as an anomalous scenario. In the case of scenarios (c) and (d), the proposed model assigns a moderately high score while one proposed by Sultani assigns a low value.

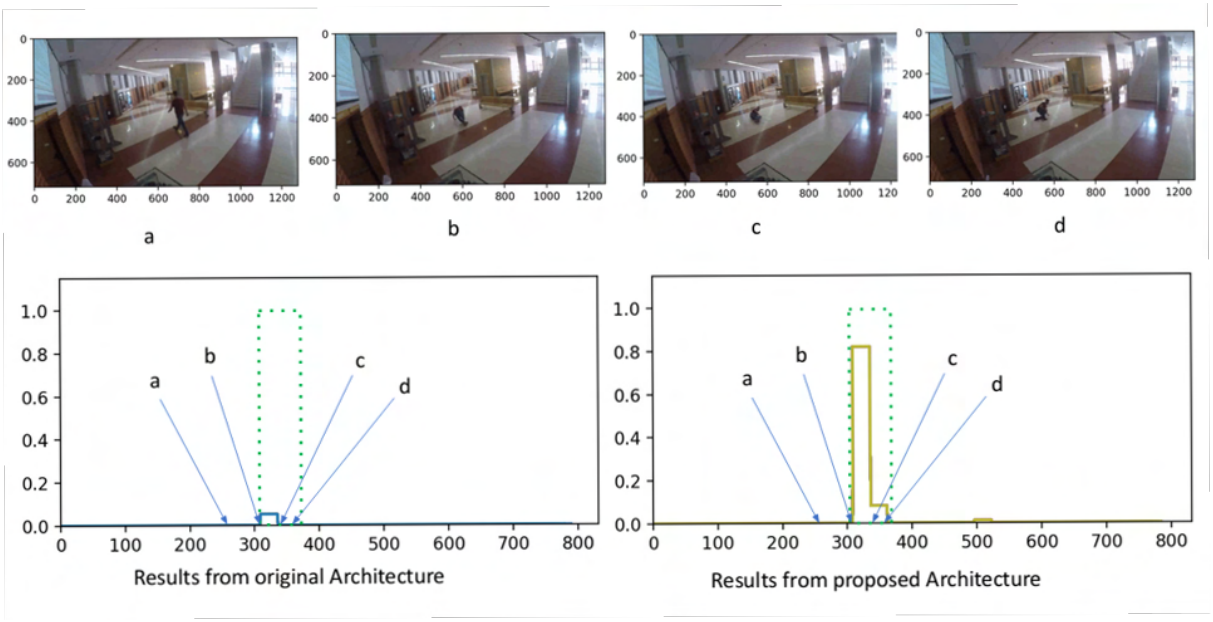


Figure 4.5: Qualitative visual results and comparison of the proposed method with [5] in far-view scenes from GBA ('Fall').

Subsequently, both models have been utilized to process a scenario from GBA dataset in near-view, showing the obtained results in figure 4.6. The second example is to the act of leaving behind objects in commonly frequented intelligent environments, such as airports, which could potentially serve as significant anomalies in the context of video surveillance.

Additionally, it encompasses four distinct circumstances. Firstly, in scenario (a), there exists a bag positioned on the ground. In scenario (b), one individual is observed retrieving the bag from the ground, while in scenario (c), another individual is seen placing a fresh bag on the ground. In scenario (d), an individual eventually directs their attention towards

the bag situated on the ground. Based on the established ground truth, scenario (a) is classified as normal, whilst the remaining scenarios are classified as abnormal.

In this particular instance, both models have produced elevated anomaly scores for scenarios (c) and (d), indicating a similar performance. In contrast, the concept outlined in this study exclusively identifies the act of retrieving a bag, hence exhibiting superior performance compared to Sultani's suggestion.

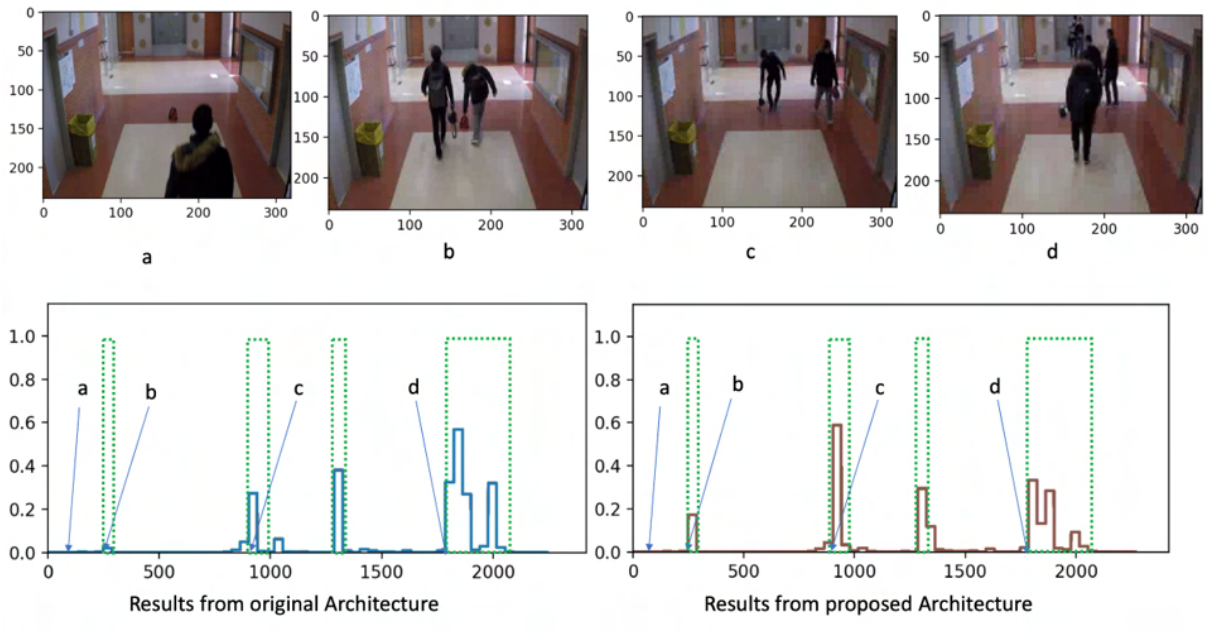


Figure 4.6: Qualitative visual results and comparison of the proposed method with [5] in near-view scenes from GBA ('Abandoned Object').

In both the scenarios of far and close examinations of the GBA dataset, the methodology put forward in this study demonstrates superior performance compared to the method Sultani *et al.* [5].

Ultimately, the users are presented with visual representations of the outcomes derived from two distinct samples extracted from the UCF-Crime dataset. This enables them to do a comparative analysis of the results obtained through both methodologies.

The first test is performed in an arrest situation, as shown in figure 4.7. In the above illustration, scenario (a) depicts a roadway featuring a stationary red automobile beside pedestrians engaged in locomotion. In scenario (b), a collision occurs between a motorbike and another car. Scenario (c) depicts law enforcement officers attempting to extract the driver from the vehicle involved in the collision, while scenario (d) portrays the culmination of their efforts as the driver is apprehended by the police. Based on the established factual information, scenario (a) can be classified as a typical occurrence, but scenarios (b, c, d) exhibit characteristics that deviate from the norm, therefore being categorized as anomalous.

When comparing the outcomes produced by the proposed model to those obtained with Sultani *et al.* in this particular case, our model demonstrates higher anomaly scores for all anomalous circumstances. This leads to the conclusion that our model performs better. Moreover, both models have similar capabilities in detecting the normal condition.

The second sample from UCF-Crime included in these results in figure 4.8 is an explosion. In this particular instance, scenario (a) portrays a typical circumstance. Scenario (b) depicts the occurrence of an explosion, whereas scenarios (c) and (d) portray the efforts of firefighters in extinguishing the fire. In the context of the ground truth, scenario (a) is classified as normal, but scenarios (b, c, d) are classified as abnormal.

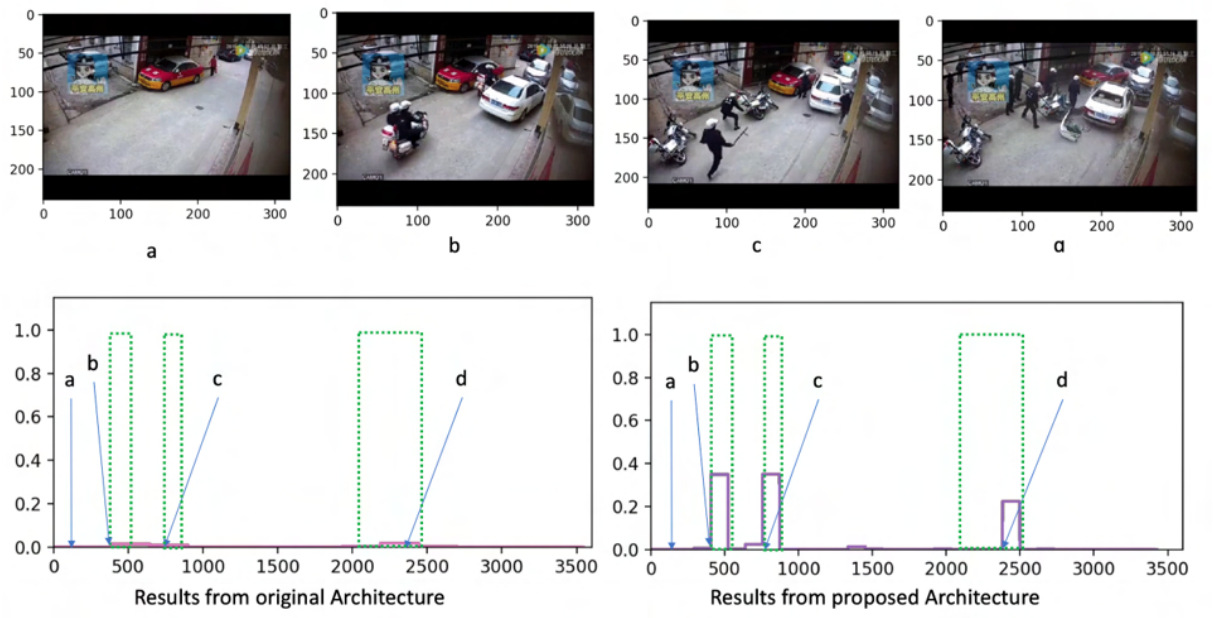


Figure 4.7: Qualitative visual results and comparison of the proposed method with Sultani *et al.* [5] in UCF-Crime ('Arrest').

Again in this particular instance, the proposed methodology demonstrates elevated anomalous scores for scenarios (b) and (c) in comparison to the scores generated by Sultani's model. In the context of scenario (d), both models exhibit comparable performance. Therefore, this particular example likewise demonstrates the exceptional performance of the suggested approach.

The comprehensive illustrations, such as those in figures 4.5 and 4.6, offer visual validations of the model's performance. In essence, the model adeptly discerns anomalies in diverse contexts, whether it's a person falling or the act of leaving behind objects in high-risk environments. These illustrations emphasize the model's superiority by highlighting its ability to produce more accurate anomaly scores compared to those obtained by Sultani *et al.*

Moreover, the UCF-Crime dataset evaluation, showcased in figures 4.7 and 4.8, further cements the efficacy of the proposed model. Scenarios ranging from arrests to explo-

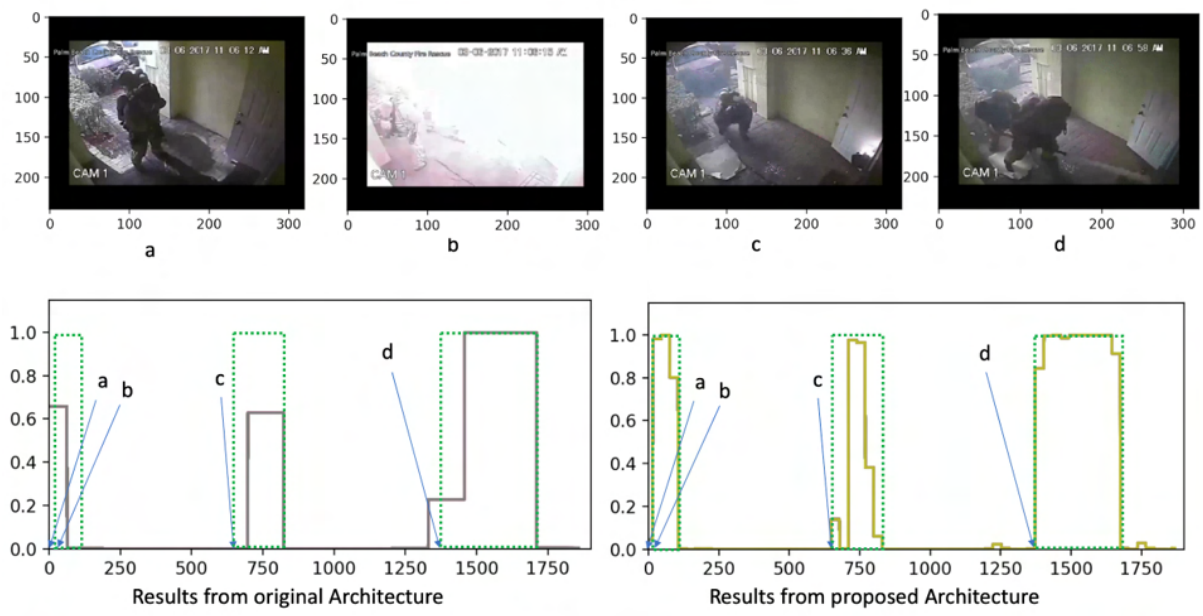


Figure 4.8: Qualitative visual results and comparison of the proposed method with Sultani *et al.* [5] in UCF-Crime ('Explosion').

sions highlight the model's ability to handle a vast array of anomalies, showcasing higher anomaly scores for most anomalous situations compared to the counterpart model.

In the grander scheme of anomaly detection in surveillance videos, the significance of such advancements cannot be understated. Surveillance applications, particularly those in high-risk environments like airports or urban centers, necessitate the utmost precision. Erroneous detections can lead to false alarms, wasted resources, or, worse, overlooked threats.

Thus, when a model demonstrates a tangible improvement over existing state of the art methodologies, as is the case here, its potential impact in real-world applications is vast. The juxtaposition with Sultani's model effectively underscores the advances made, making it a seminal contribution to the field.

4.4 Conclusions

This chapter presents a new method for finding anomalies in surveillance footage that focuses on common human actions. For feature extraction, the approach uses a T-C3D, which was previously trained using the extensive Kinetics dataset [122]. The MIL classifier is then fed with these features. Our methodology is notable for introducing a new ranking loss function that is intended to drastically lower the miss rate in anomaly categorization.

The approach outperforms cutting-edge methods, with an AUC of 80.36%. It displays improved sensitivity in anomaly identification using both the more realistic GBA dataset and the well-known UCF-Crime dataset [5], which both depict real-world circumstances for smart spaces on a regular basis.

Notwithstanding the notable outcomes, our methodology encounters difficulties when dealing with recordings characterized by rapid activities, densely populated situations, subpar visual quality, and insufficient illumination. Subsequent investigations is then needed to enhance the resilience and efficacy of the anomaly detection model in the face of challenging circumstances. One potential strategy is to integrate motion data into the model. Moreover, the focus of our research is on improving the proposed model for real-time application, aiming to develop a practical prototype. High-speed detection in video surveillance systems is essential for prompt response to anomalies, crucial for security applications.

In our upcoming work, we introduce an innovative ranking loss function integrated with an attention mechanism. This approach significantly enhances feature extraction efficiency, crucial for anomaly detection in video data. Therefore the proposed model try to outperform current weakly supervised anomaly detection methods, by means of the attention mechanism, as evidenced by extensive experiments.

We have thus shifted from traditional, exhaustive video analysis to a model that quickly pinpoints anomalies in specific scenarios, operating effectively within a few seconds,and with high reliability. This demonstrates the model's suitability for near real-time applications, as demonstrated in chapter 5 that follows.

Chapter 5

Second contribution: Real-time Weakly Supervised Anomaly Detection with Attention Mechanism

Real-time Weakly supervised Anomaly Detection with Attention Mechanism

MOHAMMAD IBRAHIM SARKER, University of Alcalá. Department of Electronics., Spain
MARTA MARRÓN-ROMERA, University of Alcalá. Department of Electronics., Spain
CRISTINA LOSADA-GUTIÉRREZ, University of Alcalá. Department of Electronics., Spain

The use of surveillance cameras for ensuring the safety of people in public gatherings, e.g., roads, intersections, banks, shopping malls, etc., is being used increasingly in recent years. To detect violence and other abnormal behaviors from these surveillance videos, the monitoring, and detection need to be intelligent and precise. Manual monitoring of these videos is tiring and prone to errors. So automatic detection of video anomaly is a very important research topic nowadays. Although many methods have been proposed, research is still ongoing for a reliable method for violence detection. Most of the earlier state-of-the-art works suggest that supervised detection is better in anomaly detection than unsupervised one. However, given the difficulty of manually annotation and training large datasets for supervised learning, weakly supervised methods provide an excellent alternate solution. Generally, in weakly supervised learning, anomaly detection is constructed as a Multiple Instance Learning (MIL) problem. The base of that technique is to build the training model from labeled bags (comprising multiple instances) rather than labeled individual instances (in the computer vision branch: videos or frames). A bag is annotated as negative if all the instances are negative, while it is marked positive if at least one instance is positive. But the problem is that the recognition of positive instances is largely biased by the dominance of negative ones. As in the real world, abnormal events are very subtle, and the difference between normal and abnormal instances is minimal. Addressing this issue, this paper proposes a weakly supervised learning algorithm with an attention mechanism for anomalous events detection. Also, a novel ranking loss function is proposed to lessen the number of false negatives in the anomaly detection task and to extend the distance between classification scores of anomalous and normal videos in the learned model. Experimental results of this proposal within UCF-Crime and Shanghai-Tech datasets show that the proposed method outperforms state-of-the-art ones. Moreover, using an attention mechanism enhances feature extraction in the generated model inference. Finally, it is demonstrated that the proposal can be run in real-time, a contribution that can greatly impact real applications in the context of interest.

CCS Concepts: • **Computing methodologies** → **Semi-supervised learning settings**; • **Computer systems organization** → **Real-time system architecture**.

Additional Key Words and Phrases: video surveillance, anomaly detection, weakly supervised learning

ACM Reference Format:
Mohammad Ibrahim Sarker, Marta Marrón-Romera, and Cristina Losada-Gutiérrez. 2023. Real-time Weakly supervised Anomaly Detection with Attention Mechanism. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The role of surveillance cameras for insuring public safety has had a huge impact in recent years, since it has allowed decreasing the security threats such as robbery, accidents and illegal activities. Thereby, although it cannot guarantee public safety, crowded places like banks, markets, railway and bus stations are kept under continuous monitoring which can result in a decrease in crime in these places.

This chapter describes an improved method for automatically spotting anomalies in surveillance footage. Because supervised learning has drawbacks and human monitoring has its limitations (as previously explained), the authors suggest a weakly supervised learning method that makes use of an attention mechanism to improve the detection of anomalous occurrences. In order to reduce false negatives and improve the ability to distinguish between typical and anomalous occurrences, a new ranking loss function is also presented. The strategy performs better than previous methods, according to tests

on the UCF-Crime and ShanghaiTech datasets. The ability to perform this technique in real-time is crucial, as it expands its usefulness in surveillance circumstances. The methodology and main results obtained with this proposal have been presented in the ESCC 2023 [25].

5.1 Introduction

Surveillance cameras are becoming more and more important in the sphere of public safety, helping to reduce risks like robberies, accidents, and illegal activities, especially in populated places like banks, markets, and transportation hubs. However, as it is repetitive, the majority of anomaly detection currently done is done manually, which is not only time-consuming but also prone to mistakes. As a result, automated systems are required for enhanced anomaly detection [23].

As it has been explained in chapter 2, for circumstances involving video surveillance, many solutions have been developed, with the earliest ones being mostly image-based [12]. Even if they are easy to understand and exact, these approaches frequently miss complex or diverse anomalous events. Due to this restriction, research into unsupervised or weakly supervised detection techniques that may identify anomalies without explicit training is now necessary. Despite their benefits, these methods have certain disadvantages as well. For example, the supervised approach requires extensive labeling whereas the unsupervised approach has high false positive rates. Thus, the scientific community is actively investigating weakly supervised classification algorithms, which strike a balance between low training requirements and tolerable accuracy levels, to address these issues.

There are many techniques for video action recognition thanks to the development of deep learning [13]. Some popular feature extraction methods include ResNet, which utilizes skip connections for effective 2D CNN structures [14], and I3D [2], which inflates the 2D convolutional filters of the Inception network for video processing, albeit at a higher computational cost [129].

This chapter describes the second contribution of the PhD thesis, which include the introduction of an improved ranking loss function with an attention mechanism and the real-time applicability of our model, underscore the innovation and practicality of our approach for weakly supervised anomaly detection.

The proposed method is validated using two benchmark ShanghaiTech dataset [4] and UCF-Crime [5], previously described in chapter 3. Experimental results reveal that our attention mechanism significantly enhances feature extraction from videos for normality classification tasks. Moreover, the application of pre-trained features from C3D [130], I3D [2], and 3D-Resnet-152 [3] within our attention mechanism outperforms the current state of the art results across all benchmarks.

5.2 Methodology

5.2.1 Proposed Architecture

In this chapter, we formulate anomaly detection, again, as a MIL problem. A general scheme of the proposed architecture for anomaly detection is shown in figure 5.1.

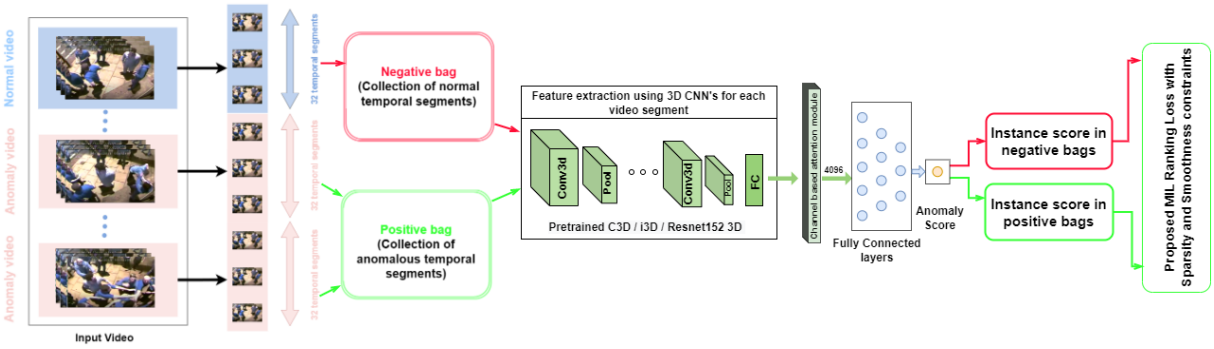


Figure 5.1: Proposed architecture for anomaly detection with attention mechanism in surveillance videos.

In our research, as depicted in figure 5.1, we addressed the challenge of surveillance video analysis through a structured partitioning process, wherein each video is divided into 32 distinct, non-overlapping temporal segments. This segmentation organizes the video data into manageable instances for our weakly supervised learning framework.

Following a methodology inspired by our 1^{st} Contribution [23] and the one in [5], for each of these segments, we delve deeper into the temporal dynamics by processing subclips, now consisting of 30 frames each. The increase in frames per subclip from 16, proposed in [5], to 30, here used, allows for a more comprehensive capture of both spatial arrangements and temporal movements within each segment, pivotal for distinguishing anomalous behavior from normal activities.

Additionally, we introduced a learnable aggregation weight [131], functioning as an attention mechanism, to instill a degree of regularity within each bag, corresponding to the 32 temporal segments.

A fundamental aspect of our work is the extraction of video-temporal and visual features through the use of 3D CNNs. Notably, we leveraged various network architectures such as C3D [130], I3D [2], and 3D-Resnet-152 [3] as feature extractors, enabling result comparison based on the AUC metric.

As part of our preprocessing procedure, we resized each video frame according to the specific network in use: 240x320 pixels for both C3D and 3D-Resnet-152, and 224x224 pixels for I3D. The feature extraction process involved segmenting the input video into sequences of 16 frames, which were then fed into the CNN. Subsequently, we calculated an average across all the features from these 16 frames, originating from the Fully Connected (FC) layer output. This step was pivotal in obtaining a compressed feature with a size of

4096, a choice influenced by typical deep learning network architectures, where layer sizes are often powers of 2, balancing computational efficiency and representational power [132].

The feature extraction process preserved both spatial and temporal correlations. The introduction of selection mechanisms enhanced the feature's discernment, ultimately contributing to improved performance. In the final stages, a multi-headed channel-based attention mechanism was applied to each generated feature, identifying the salient components crucial for the subsequent classification.

The attention layer, maintaining its original size of 4096, was then fed into a fully connected layer, where the classification into anomalous or normal segments was performed. This comprehensive approach forms the basis of our research methodology.

5.2.2 Attention Mechanism

As mentioned in section 5.1, temporal instances of anomalous or atypical events have random nature, and thus conventional classification techniques cannot model these anomaly patterns. Hence, anomaly detection in video segments is formulated in the form of a MIL model, as most videos of the available datasets are annotated on a video-level fashion instead of a segment-level one. Typically, in such datasets, those containing anomalies are considered as **'positive'**, whereas the rest are **'negative'**.

Moreover, a surveillance video annotated as anomalous (i.e. a **'positive video'**) represents a **'positive bag'**, where each of its temporal segments is regarded as a separate **'positive instance'**. Likewise, a normal video (i.e. **'negative video'**) represents a **'negative bag'**, which is also divided into instances. Usually, the training models are based on a predetermined number of instances also named here as **'segments'** [23].

In typical MIL models, attention blocks are then utilized to add context within each video into the classification. Therefore, each segment in a video can be differentiated as normal or anomalous by assigning them the appropriate attention weight. The attention mechanism is thus added to enhance the feature extraction in the classification task.

Let $\mathcal{B} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a bag of n features extracted from the proposed network output, where $\mathbf{x}_i$ represents the i-th feature vector. The function $\phi(\mathbf{x}_i)$ is applied to each feature vector to enhance its representational power, crucial for the attention mechanism. We then represent the attention function a_n as in Equation 5.1, where each transformed feature vector $\phi(\mathbf{x}_i)$ contributes equally (weighted by $\frac{1}{n}$) to the formation of the comprehensive feature vector $\mathbf{h}$, though the actual weights are dynamically calculated based on the attention mechanism.

$$h = \sum_{n=1}^{N} a_n \phi(\mathbf{x})_n$$
$$a_n = \frac{\exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{A}\phi(\mathbf{x})_n^\top\right)\right\}}{\sum_{j=1}^{N} \exp\left\{\mathbf{w}^\top \tanh\left(\mathbf{A}\phi(\mathbf{x})_\mathbf{j}^\top\right)\right\}}, \tag{5.1}$$

where $\mathbf{w} \in \mathbb{R}^{N\times 1}$ and $\mathbf{A} \in \mathbb{R}^{N\times M}$ are the attention mechanism adjustable parameters.

Furthermore, we use the non-linearity of *tanh* function (i.e., a hyperbolic tangential function) on an element-to-element basis to incorporate positive and negative values in the attention weight a_n to ensure appropriate flow of gradient. This formulation is imperative to explore the existence of anomalies (or lack of them) within the different segments of the surveyed video. Thus, the output h is then trained and used later for the classification model.

5.2.3 Negative Ranking Loss Function

The primary distinction between the proposal described in chapter 4 ([23]) and the method proposed here lies in their respective loss functions and the underlying methodologies for video anomaly detection.

In [23], the central focus is on MIL, which treats videos as collections of segments and aims to effectively identify segments with anomalous behavior. The key to [23]'s approach is the loss function, which seeks to optimize the model's parameters by minimizing the negative log likelihood of the predicted anomaly scores, signifying the likelihood that a segment contains an anomaly. It achieves this by aligning the model's predictions with the actual labels, striving to bring the predicted scores into closer agreement with the actual anomalies. Essentially, this loss function encourages the model to be more precise in pinpointing the temporal segments where anomalies occur.

In contrast, the proposed method, while also rooted in MIL, formulates the problem as a ranking task. It endeavors to not only detect anomalous segments but also rank them by their degree of anomaly. This ranking loss function allows for a more nuanced assessment of segment anomalies and aligns with the intricacies of real-world surveillance scenarios, where anomalies may vary in severity and distribution. Furthermore, the proposed method is distinguished by its use of benchmark datasets and the frame-level AUC metric for evaluation, underscoring its practical and specialized contributions to the field of video anomaly detection. The ranking loss in the proposed method is meticulously designed to capture temporal context, encouraging the model to identify and rank the most anomalous segments within each video. Additionally, the attention mechanism plays a critical role in feature extraction and context addition, emphasizing the holistic view of video anomalies.

Moreover, the proposed method formulates the problem as a negative likelihood problem, aiming to capture the temporal context within videos. This approach differentiates

between the various anomalies that may appear in a single surveillance footage. The proposed loss function incorporates sparsity limits and a multi-instance attention strategy, enhancing feature extraction, addressing training error, and handling information loss. This binary classification for segmenting abnormalities in videos is addressed by the attention modeling, resulting in a comprehensive and advanced approach to video anomaly detection.

To achieve this, the proposed method formulates equation 5.1 as a negative likelihood problem, as demonstrated in equation 5.2. This formulation is critical in encouraging the model to differentiate between normal and anomalous segments, maximizing the alignment between model predictions and ground truth labels while incorporating regularization terms to ensure model robustness.

$$\max_{\mathbf{w}} \left[\frac{1}{z} \sum_{j=1}^{z} \sum_{i \in \mathcal{B}_j} y_i \log \left(\mathbf{w} \cdot h_i - b \right) + \left(1 - y_i \right) \log \left(1 - \left(\mathbf{w} \cdot h_i - b \right) \right) \right] + \|\mathbf{w}\|^2 \tag{5.2}$$

In the given equation, w represents the weight vector of the model that is being optimized, z denotes the total number of bags or groups of video segments, and $\mathcal{B}_j$ is the j-th bag containing multiple segments. Each segment within a bag is labeled with y_i, indicating whether it's an anomaly (1) or normal (0). The feature vector h_i of each segment is obtained after processing through a neural network or similar transformations to capture essential features. The term b serves as the bias term. The natural logarithm, denoted by log, is utilized in computing the log-likelihood, which is crucial for estimating the probabilities in logistic regression frameworks. Finally, $\|\mathbf{w}\|^2$ is a regularization term, likely $L2$ regularization, which helps prevent the model from overfitting by penalizing the magnitude of the weights, ensuring that the model maintains generalizability across different datasets.

With this loss function, multi-class classification can be achieved, as the negative log function ($-log(y)$) provides a prediction related to each video segment. Subsequently, the overall loss for each bag is calculated by averaging all of the contained segments (i.e. the mean of each bag of segments).

In this scenario, the anomaly detection in videos is represented as an anomaly value regression. The obtained model allows the videos with surveillance anomalies (**'positive videos'**) to get higher scores compared to the videos without any anomaly (**'negative videos'**).

Besides, if there are annotations on the segment level, the ranking loss can be computed as follows in equation 5.3.

$$f(Va) > f(Vn), \tag{5.3}$$

where f represents a mapping function to correlate any segment with its respective predicted scores, normalized between 0 and 1.

In fact, Vn and Va in the above inequality mean normal and anomalous segments, respectively. Typically, this function f can be implemented as a Fully Connected Neural Network (FCNN). Nonetheless, in the case of interest, annotations are limited to video-level $\mathcal{V}$, and not segment-level V, as explained before. Therefore, the ranking loss function in equation 5.3 has to be formulated as follows in equation 5.4.

$$\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) > \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right), \tag{5.4}$$

where, max function is used to obtain the maximum value for all segments in each bag of the video.

The rationale behind this formulation is that segments in the **'positive bag'** ($\mathcal{B}_a$) with the highest anomaly scores must have bigger importance if compared to that of segments in **'negative bags'** ($\mathcal{B}_n$), that represents no anomaly at all.

Consequently, the resulting ranking loss function $l\left(\mathcal{B}_a, \mathcal{B}_n\right)$ is represented as shown in equation 5.5.

$$l\left(\mathcal{B}_a, \mathcal{B}_n\right) = -\log\left(\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) \tag{5.5}$$

In this negative loss function, there are two problems. One, it overlooks the temporal information of any surveyed video with anomalies: What if a video consists of several segments with anomalies? Therefore, it is imperative to include a temporal context to distinguish between the different anomalies that can appear in a single surveillance video. The other issue is that this loss function can lead to a prediction where the majority of segments are returned as normal.

To address the limitations, constraints for giving smoothness (with λ_2) and sparsity (with λ_1) to each video instance score are included in the loss function as described in equation 5.6.

$$l\left(\mathcal{B}_a, \mathcal{B}_n\right) = -\log\left(\max_{i \in \mathcal{B}_a} f\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} f\left(\mathcal{V}_n^i\right)\right) + \lambda_1 \sum_{j=i}^{n-1}\left(f\left(\mathcal{V}_a^j\right) - f\left(\mathcal{V}_a^{j+1}\right)\right)^2 + \lambda_2 \sum_{j=i}^{n} f\left(\mathcal{V}_a^j\right) \tag{5.6}$$

Since the occurrence of anomalies is extremely rare in the video surveillance task, it is imperative to use constraints for sparsity. Only a few video segments might contain high scores for anomalies in positive videos. Finally, we obtain the global model loss function $\mathcal{L}(\mathcal{W})$ as in equation 5.7 after adding sparsity constraints.

$$\mathcal{L}(\mathcal{W}) = l\left(\mathcal{B}_a, \mathcal{B}_n\right) + \|\mathcal{W}\|_F, \tag{5.7}$$

where, $\|\mathcal{W}\|_F$ describes the loss weight after applying the constraint for sparsity.

The loss function, when combined with the attention mechanism improves how the model extracts features from the video data. As the video information flows through the network, the attention mechanism steps in to recognize and allocate suitable weights to the various segments in each bag of data. These weights signify the importance of each segment in the broader classification task, essentially directing the model's attention to the most crucial temporal details. These weights serve as indicators of how important each segment is within the overall classification task, essentially guiding the model to pay special attention to the most critical temporal details.

5.3 Experimental Results and Discussion

5.3.1 Experimental Setup

Two multi-scene benchmark datasets are used for evaluating the proposal, as they were created for weakly supervised video anomaly detection. These are ShanghaiTech [4] and UCF-Crime [5], previously explained in chapter 3.

As mentioned in previous paragraphs, computational cost is one of the challenges of such models, so here, as well as reliability and functionality results, computational cost ones are also included. Our experiments were carried out on a high-performance computing platform, featuring an AMD Ryzen 9 5900X 12-core processor clocked at 3.70 GHz, 32GB of RAM, and a Nvidia RTX 3080 Graphics Processing Unit (GPU).

5.3.2 Performance Evaluation

Table 5.1 shows frame-level AUC results on the proposal in the ShanghaiTech (column 4) and UCF-Crime (column 5) datasets. As stated in the table, for the ShanghaiTech dataset, when compared with the results obtained with unsupervised learning methods [4] and weakly supervised methods Sultani *et al.* ([5]), the proposed one achieves better results than the rest state of the art methods. Moreover, when using I3D features and the proposed attention mechanism, our model achieves an accuracy of 91.21%.

Analyzing the results for the UCF-Crime dataset (column 5 in table 5.1, it can be noticed that the proposed method outperforms all previous unsupervised learning approaches. Remarkably, using I3D-Attention and 3D-ResNet-152 Attention mechanisms, the proposed method outperforms the current state of the art MIL-based methods, as the accuracy in above-mentioned methods are 83.70% and 87.66% respectively (see [3]).

Table 5.1: Comparison of frame-level AUC performance with other state of the art unsupervised and weakly supervised methods on ShanghaiTech (column 4) and UCF-Crime (column 5).

Supervision	Method		Feature	AUC (%)	
				ShanghaiTech	UCF-Crime
Unsupervised	Stacked-RNN	[110]	-	68.00	-
	FFP	[4]	-	73.4	-
	Mem-AE	[113]	-	71.20	-
	VEC	[119]	-	74.80	-
	Lu *et al.*	[109]	C3D-RGB	-	65.51
	GODS	[15]	I3D-RGB	-	70.46
Weakly Supervised	GCN	[117]	C3D-RGB	76.44	81.08
			TSN-Flow	84.13	78.08
			TSN-RGB	84.44	82.12
	Zhang *et al.*	[116]	I3D-RGB	82.50	-
			C3D-RGB	-	78.66
	Sultani *et al.*	[5]	C3D	-	75.41
	Motion Aware	[9]	PWC Flow	-	79.00
	1^{st} **Contribution**	[23]	T-C3D-$\log_2$	-	80.36
	Our proposal	[25]	I3D-Att.	**91.21**	**83.69**
			3D-ResNet-152 Att.	-	**87.65**

5.3.3 Computational Cost

In table 5.2, we summarise and compare frame rates of different state of the art methods with the UCF-Crime dataset. In order to conclude and rate the performance of the proposed algorithm on real-time monitoring scenarios, we devised a simulation system to visualize the model prediction.

Table 5.2: Processing time per frame in ms.

Method		Processing time (ms/frame)
Spatio-temporal Texture Modelling	[56]	18.4
Incremental Spatio-temporal Learner	[112]	37.0
Non parametric modeling	[108]	125.0
Holistic Features	[55]	25.0
VEV	[33]	7.0
Our proposal		**5.0**

For real-time anomaly detection, videos are temporally segmented into batches of 16 consecutive frames. These frames are processed by a 3D-ResNet-152 feature extractor (We have also experimented with I3D feature extractors, finding that the processing times are nearly identical), which generates the necessary features to feed into the anomaly detection network. The output of the proposed model is then always a classification score between 0 and 1, as described by equation 5.7, signifying with 0 that the event surveyed in the input video is normal, and 1 that the event is anomalous, and taking into account that it

is generated an output per video frame. In such situations, the average time to process one frame with the global proposed flowchart around 5ms.

5.3.4 Qualitative Results

Qualitative results of our proposal are presented, including sample videos from both the UCF-Crime and ShanghaiTech datasets.

In figures 5.2 and 5.3, we showcase segments from the UCF-Crime dataset. Figure 5.2 exemplifies a normal class scenario, depicting a typical scene within a supermarket. In figure 5.3, we delve into an abnormal class, where an explosion occurs, distinctly marked by a conspicuous blue spike. The green marked area in these frames corresponds to the ground truth.

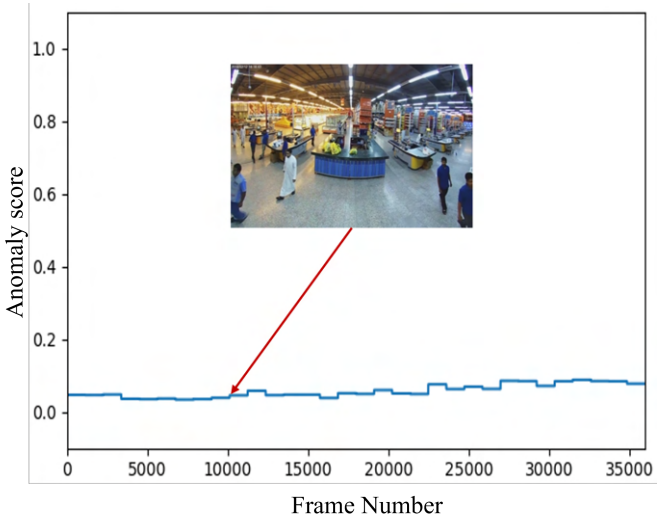


Figure 5.2: Qualitative visual results for UCF-Crime dataset: normal class.

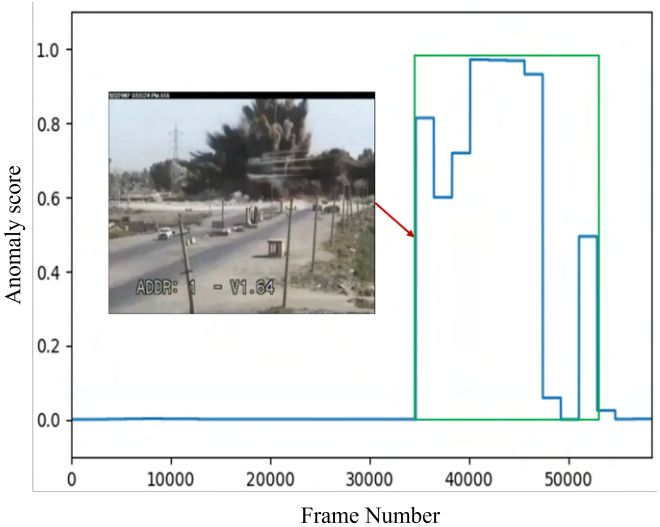


Figure 5.3: Qualitative visual results for UCF-Crime dataset: anomalous class.

Moving on to the ShanghaiTech dataset, figure 5.4 captures a normal class scenario where two students are leisurely walking along a road. This represents a common and expected scene. However, in figure 5.5, we introduce an anomaly scenario, depicting a student walking on the road while skillfully spinning an object above their head. This unusual and intriguing behavior contrasts with the normal class scenarios in the dataset.

Furthermore, for each video frame, we have plotted a final classification parameter, a value ranging from 0 to 1. This parameter illustrates higher peaks for the anomalous segments within the video, offering a clear visualization of the detection and classification of anomalous events in the presented datasets.

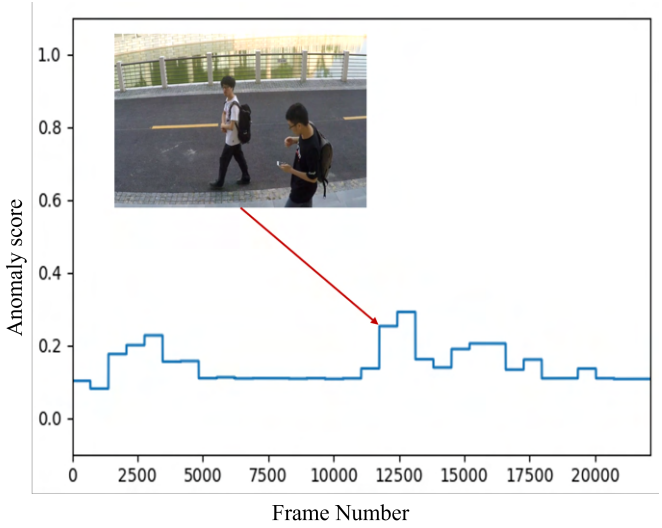


Figure 5.4: Qualitative visual results for ShanghaiTech dataset: normal class.

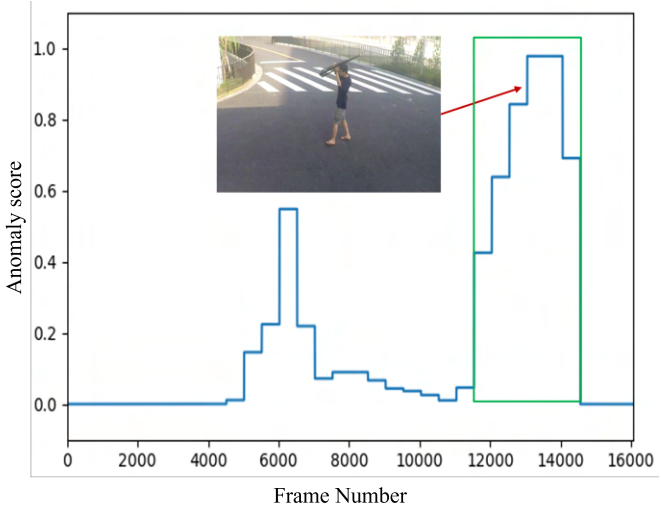


Figure 5.5: Qualitative visual results for ShanghaiTech dataset: anomalous class.

5.4 Conclusions

In this chapter, an improved weakly-supervised method for finding abnormalities in surveillance video data is presented. Our approach sets itself apart by a number of significant contributions:

- **Improved Ranking Loss Function with Attention Mechanism.** an improved ranking loss function is introduced in the suggested model and is combined with an attention mechanism. This improved combination makes it possible for feature extraction, which is essential for finding anomalies in video data, to be more effective and efficient.
- **Superior Performance.** Our model performs better when compared to current state of the art techniques for weakly supervised anomaly identification. Results from many experiments, both quantitative and qualitative, serve as proof of this.
- **Real-Time Applicability.** We have used real-time scenarios to validate the suggested architecture. The model is capable of real-time applications since it operates quickly and efficiently detects anomalies.

This study highlights the effectiveness of weakly supervised learning techniques for anomaly identification. It also emphasises the need for further developments and additions to the suggested model.

Moving forward, we want to make our anomaly detection system more reliable by combining several data sources, like audio, video, and sensor data. According to our estimates, doing so would give a more complete representation of the surveillance environment and enhance the system's capacity to spot anomalies.

Our research also emphasises the significance of frame-level labeling in weakly-supervised learning techniques. Our research has shown that labeling frames at the frame level can considerably enhance the learning of spatio-temporal properties. In order to improve the performance of weakly-supervised anomaly detection systems, we plan to further investigate this insight in next research.

As a result, our study opens the door for additional investigation and advancements in the area of weakly-supervised anomaly identification.

Chapter 6

Advancements in Deep Learning for Real-Time Video Anomaly Detection

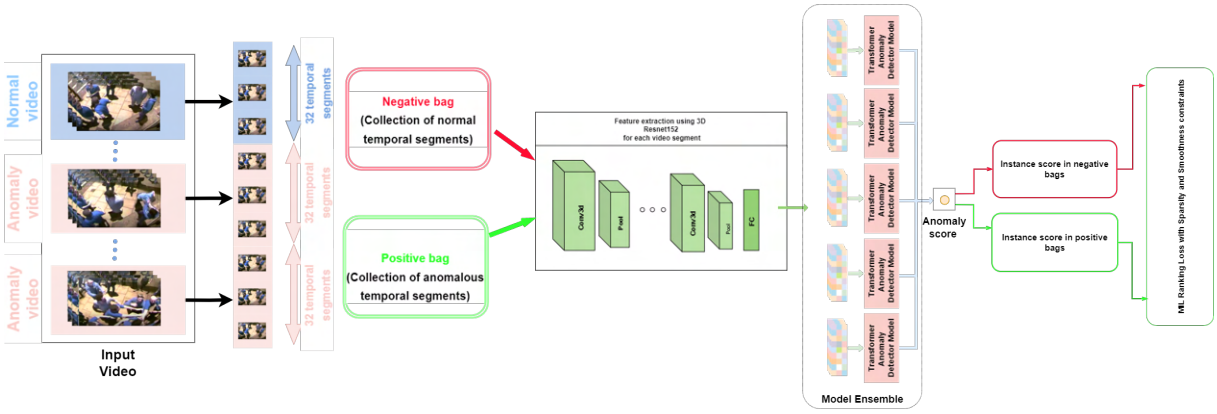


Figure 6.1: Final proposal for anomaly detection in surveillance videos.

This chapter describes in detail the complete proposal for anomaly detection developed in this PhD. The different parts that conform the proposal with its importance and contributions are detailed in this chapter.

6.1 Introduction

As it has been detailed in previous chapters, anomaly detection in video surveillance is a critical component of modern security systems, playing a vital role in identifying unusual or suspicious activities in real-time. The ability to promptly detect and respond to these anomalies is essential for ensuring public safety and protecting assets. Building upon previous research utilizing DL techniques, this chapter presents an improved and sophisticated approach to VAD, aimed at addressing the inherent challenges of real-time video surveillance.

Figure 6.1 presents a general block diagram of our proposed VAD system. This system starts with the crucial step of pre-processing video sequences. Each video frame is resized to a common resolution of 112×112 pixels. This resizing is a key step in balancing computational demands with the retention of important visual information, which is essential for effective anomaly detection.

Following the resizing, the video is divided into segments of 32 temporally consistent frames. This segmentation technique, detailed in section 4.1, ensures that the temporal context is maintained, which is critical for identifying anomalies over time.

In our approach, we employ the 3D-ResNet-152 model for feature extraction, a decision based on its standout performance in our initial experiments (detailed in chapter 5). This model, compared to others such as T-C3D, C3D, and I3D, excels in extracting a wide range of features. Its design, including residual connections, helps in efficient model training as well as in overcoming the vanishing gradient problem [133]. Moreover, the 3D layers in this model are particularly effective in capturing temporal dynamics in video sequences.

To enhance the capability of the 3D-ResNet-152 model, we integrate it with the MIL framework. This integration enables the model to focus more effectively on the most significant information, thereby enhancing its performance in understanding the spatio-temporal nuances in videos. The output from this process is a feature vector of dimensions 1×4096 for each temporal batch, providing a condensed but informative representation of each video segment.

To further improve the ability to perceive temporal dynamics, we use an ensemble of models, including the Attention 3D-ResNet-152, Transformer 3D-ResNet-152, and an Ensemble of 3D-ResNet-152 Transformer models. The characteristics of these models are elucidated in section 6.4.

A cornerstone of our Transformer-based models is the Transformer encoder [24], whose architecture is depicted in figure 6.2. Utilizing feed-forward networks and self-attention processes, this encoder discerns the importance of frames and understands lengthy temporal sequences. Such capabilities are vital for identifying anomalies that persist or evolve over time.

This introduction aims to set the stage for a detailed exploration of our innovative approach to VAD. The subsequent sections will delve into the input video-sequence pre-processing and feature extraction (section 6.2), the specialized loss function used for training the anomaly detection system (section 6.3), the distinct elements of our model architecture (section 6.4), their training procedure (section 6.5) and the implementation of real-time anomaly detection in practical applications, culminating in section 6.6.

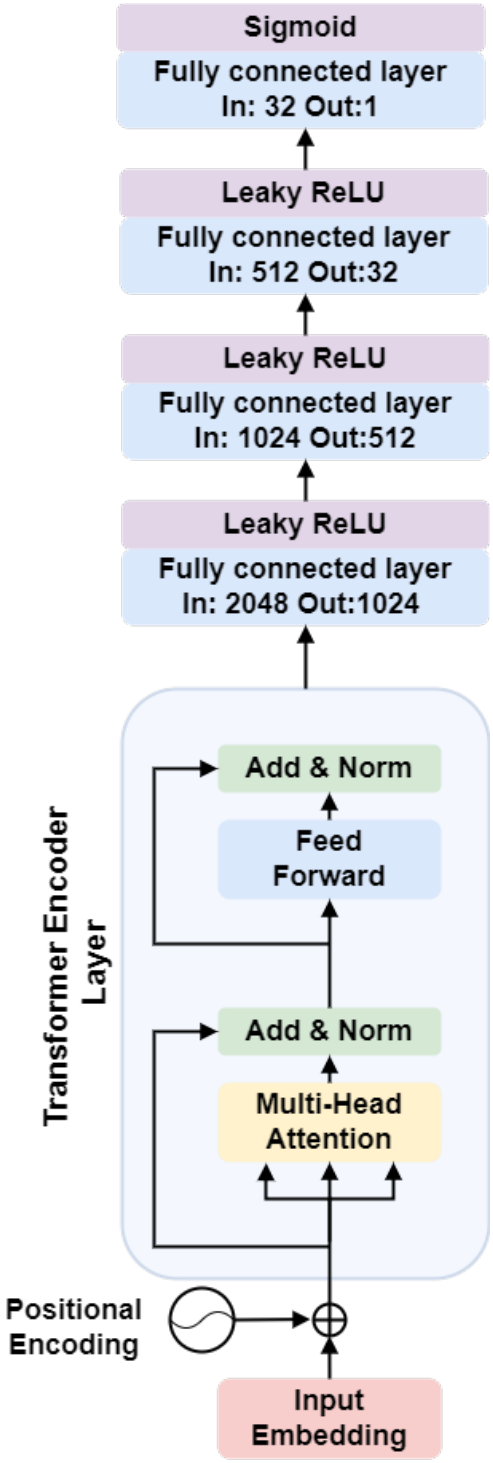


Figure 6.2: Transformer encoding structure.

6.2 Video Preprocessing and Feature Extraction

The foundation of our VAD system lies in the meticulous preprocessing of video sequences and the extraction of significant features. This section elucidates the methodologies employed in these initial stages.

6.2.1 Video Preprocessing

The foundation of our VAD system lies in the meticulous preprocessing of video sequences and the extraction of significant features.

As outlined in section 6.1, our approach starts with the input of a video sequence. Each frame of this sequence is resized to a standard resolution of 112×112 pixels. This resizing is crucial for striking a balance between computational demands and the retention of essential visual information.

Following the resizing, the video data undergoes segmentation into batches, each comprising 32 temporally consistent frames. This segmentation, detailed in section 4.1, provides a meaningful temporal context while maintaining computational efficiency. Each segmented batch is structured as a 4-dimensional matrix with dimensions $32 \times 112 \times 112 \times 3$, which includes:

- **32 Frames per Batch:** representing the temporal context for effective analysis.
- **112 x 112 Spatial Dimensions:** ensuring uniformity in the data processed.
- **3 RGB Color Channels:** providing a comprehensive color-based representation for each pixel.

6.2.2 Feature Extraction with 3D-ResNet-152

The 3D-ResNet-152 model, mentioned in the publication by Hara et al. [134] and specifically designed for analyzing spatio-temporal data, is utilized to extract features. This model takes in a matrix of size $32 \times 112 \times 112 \times 3$ from each video segment, and applies convolutions and pooling operations across both the spatial and temporal dimensions. The approach accurately identifies important spatial and temporal characteristics that are essential for detecting anomalies.

The 3D-ResNet-152 produces a one-dimensional feature vector with dimensions of 1×4096 for each temporal batch. Although it is small, this feature vector contains the fundamental patterns and subtle details from the original set of frames, which are crucial for detecting abnormal behavior in video streams.

By combining these preprocessing and feature extraction approaches, our VAD system is enhanced with superior, relevant data. The homogeneity of the input data, together with the advanced capabilities of the 3D-ResNet-152 model, establishes a basis for precisely detecting abnormalities in video sequences. It is worth noting that the representation plays a vital role in the following anomaly identification process, as it encompasses both the spatial and temporal characteristics of the video data.

6.2.3 Integration of MIL with 3D-ResNet-152

A key aspect of our architecture is the integration of the MIL technique with the 3D-ResNet-152 model. The use of MIL allows for the classification of groups, or 'bags', of instances rather than individual instances. This approach is particularly well-suited for video data, as it can effectively detect anomalies that span multiple frames. For a detailed discussion on the concept of MIL and 'bags' as used in this context, refer to chapter 4.

Every bag within our system corresponds to a distinct section of video frames. The 3D-ResNet-152 model effectively prioritizes the most important elements inside the segments by employing the MIL methodology. This strategy enhances the model's ability to differentiate between normal and abnormal patterns in the video. It achieves this by guiding the model's attention towards the most informative features within each group. Thus, integrating MIL into the 3D-ResNet-152 model significantly improves the overall accuracy and efficiency of the system in detecting anomalies.

6.3 Advancements in Video Anomaly Detection: A Weakly Supervised Transformer Model Approach

As it has been explained in section 2.3, the weakly supervised learning framework is a prevalent strategy for anomaly identification in video surveillance, being predominantly employed the MIL.

Labels in this scenario are associated with groups of instances, known as collections or 'bags', rather than with individual instances. This arrangement enables us to operate autonomously. A bag that exclusively contains negative instances is classed as negative, representing regular segments. On the other hand, a bag that includes positive examples is labeled as positive, indicating an anomalous segment.

Further developing upon the ideas discussed in our initial research in chapter 4, we emphasize the importance of the weakly supervised learning framework in the field of video surveillance anomaly detection. The MIL technique is crucial for managing labels associated with collections or 'bags' of instances. This approach enables a more comprehensive and efficient examination and is consistent with the approaches developed in our original research.

Let's further explain by considering a collection of bags $\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_n$. Each bag $\mathcal{B}_j$ consists of a group of segments $h_{j1}, h_{j2}, ..., h_{jm}$ and is assigned a label y_j. The value of the label y_j is 1 when anomalies are present and 0 when they are not. Hence, the goal is to acquire a model that can effectively distinguish between regular and anomalous segments.

In the context of MIL, a common loss function is the ranking loss. We define the scoring function as $f(h) = w \cdot h - b$, where w represents a weight vector, and b signifies

a bias term, as described in chapter 5. The ranking loss can be then articulated as in equation 6.1:

$$L_{rank}(w,b) = \sum_{j=1}^{n} \left[y_j \cdot \max_{h \in B_j} f(h) + (1 - y_j) \cdot \min_{h \in B_j} f(h) \right] \tag{6.1}$$

The aim is to minimize the ranking loss while determining the weight vector w and the bias term b. This ensures that the model allocates higher scores for anomalous segments and lower ones for normal segments. It is worth noting that the gradient of the loss function with respect to w and b can be computed using the following equations:

$$\frac{\partial L_{rank}(w,b)}{\partial w} = \sum_{j=1}^{n} \left[y_j \cdot \frac{\partial \max_{h \in B_j} f(h)}{\partial w} + (1 - y_j) \cdot \frac{\partial \min_{h \in B_j} f(h)}{\partial w} \right] \tag{6.2}$$

$$\frac{\partial L_{rank}(w,b)}{\partial b} = \sum_{j=1}^{n} \left[y_j \cdot \frac{\partial \max_{h \in B_j} f(h)}{\partial b} + (1 - y_j) \cdot \frac{\partial \min_{h \in B_j} f(h)}{\partial b} \right] \tag{6.3}$$

However, the conventional ranking loss might fall short in effectively distinguishing between instances that are reliably and correctly classified. To address this shortcoming, we propose a modification to the base loss function, as in equation 6.4:

$$L_{new}(w,b) = L_{rank}(w,b) + \lambda \cdot \sum_{j=1}^{n} \sum_{h \in B_j} [g(f(h))] \tag{6.4}$$

In the above equation, $g(f(h))$ is a novel component that encourages the dispersion of response scores within the unit interval. Meanwhile, λ acts as a regularization parameter that governs the trade-off between the ranking loss and the new term. $g(f(h))$ can be crafted to capture various properties of the response scores, such as sparsity or smoothness, based on the specific requirements of the problem. Importantly, $f(h)$ represents the output of a Transformer encoder, which processes the input data h into a feature representation that captures complex patterns relevant for anomaly detection. This highlights how the model leverages the Transformer's capability to handle long-range dependencies and contextual relationships within the data.

Utilizing this weakly supervised learning framework and optimizing the augmented loss function $L_{new}(w,b)$ allows us to train a VAD model. This model exhibits enhanced efficiency in discriminating between normal and anomalous segments in video surveillance data, thus providing a significant improvement over existing approaches.

In this context, the proposed method uses a unique loss function. The $-log$ term in our previous contribution (see chapter 4) is employed within a ranking loss function specifically tailored for anomaly detection in videos. This function focuses on accurately ranking and identifying anomalous segments or **'bags'**, optimizing for the correct identification of these anomalies. In contrast, our proposed $-log$ function is used within a more complex

loss function that accounts for both normal and anomalous segments. This modified loss function is designed to not only optimize the model's predictions by ensuring that anomalous segments are scored higher than normal ones but also to handle complex temporal dependencies introduced by the integration of Transformer models. These Transformers enhance the model's ability to understand and predict temporal patterns, thereby improving discrimination between normal and anomalous activities over time. To guide the learning process of the Transformer model, this function is detailed in equation 6.5, which effectively leverages the $-log$ component to manage a broader range of video segment classifications.

$$\max \mathbf{w}, T \left[\frac{1}{z} \sum_{j=1}^{z} \sum_{i \in \mathcal{B}_j} y_i \log \left(\mathbf{w} \cdot T(x_i) - b \right) + (1 - y_i) \log \left(1 - (\mathbf{w} \cdot T(x_i) - b) \right) \right] + \|\mathbf{w}\|^2 \tag{6.5}$$

In the above equation, the model aims to find the optimal weight values $\mathbf{w}$ that minimize the negative log likelihood function. This adjustment is made to prevent overfitting, as minimizing the negative log likelihood aligns with the typical objective in machine learning. The Transformer Model parameter T_i with inputs x_i represents a transformation function applied to the input data x_i, converting it into a feature-rich format that is more suitable for analysis by the learning model.

The objective function consists of two main components, the log likelihood term and an $L2$ regularization term $\|\mathbf{w}\|^2$. The log likelihood term computes the difference between the predicted and actual labels for each video segment, providing a probabilistic interpretation of the predictions. The $L2$ regularization term is added to control the magnitude of the weight values, penalizing large weights and promoting model simplicity. Together, these components aim to strike a balance between fitting the training data well and preventing overfitting.

However, this basic formulation of the loss function lacks consideration for the temporal context in the video data. To incorporate this important aspect, we introduce a ranking loss function (equation 6.6). Here, the highest scoring segments in the **'positive'** and **'negative'** bags are compared. As explained in previous chapters, a **'positive bag'** refers to a bag that contains at least one anomalous segment, while a **'negative bag'** refers to a bag composed solely of normal segments.

Further refinement of the ranking loss function accounts for sparsity and smoothness in the video data, as shown in equation 6.6:

$$\begin{aligned} l(\mathcal{B}_a, \mathcal{B}n) = &- \log \left(\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max i \in \mathcal{B}_n f(\mathcal{V}_n^i) \right) \\ &+ \lambda_1 \sum_{i=1}^{n-1} \left| f(\mathcal{V}_a^i) - f(\mathcal{V}_n^i) \right| + \lambda_2 \sum_{i=1}^{n-1} \left(f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}) \right)^2 \end{aligned} \tag{6.6}$$

After adding these constraints to the loss function, we obtain the final global model loss function $\mathcal{L}(\mathcal{W})$ as shown in equation 6.7:

$$\mathcal{L}(\mathcal{W}) = l\left(\mathcal{B}_a, \mathcal{B}_n\right) + \|\mathcal{W}\|_F, \tag{6.7}$$

The Negative Ranking Loss Function, which considers the temporal context of video data, compares the highest scoring segments in the positive and negative bags. The additional smoothness and sparsity restrictions enhance the performance of the loss function even further, enabling accurate anomaly identification and reducing false positives. Through the use of Transformer models and the Negative Ranking Loss Function, our recommended method significantly improves the performance of VAD.

In conclusion, by fusing the power of Transformer models with a special loss function, the Negative Ranking Loss Function, our solution overcomes the challenges posed by VAD. Because it integrates smoothness and sparsity constraints and takes into account temporal dynamics, our method is more effective at identifying anomalies in video data.

6.4 Anomaly Score Generation and Ensemble Approach

In this section, we delve into advanced architectures that have been harnessed to capture the intricate spatial and temporal patterns in video data. These architectures, based on DL techniques, are tailored to enhance the sensitivity and specificity of anomaly detection in complex video streams. Specifically, we discuss three primary architectures:

1. The **Attention 3D-ResNet-152 (A3DR)**, which enhances feature relevance.
2. The **Transformer 3D-ResNet-152 (T3DR)**, designed towards capturing long-range dependencies.
3. The **Ensemble of 3D-ResNet-152 Transformer Models (E3DRT)**, which amalgamates collective knowledge for robust detection.

With 3D-ResNet-152 [134] models, we take advantage of attention processes and ensemble learning to boost feature relevance, capture long-range dependencies, and boost anomaly detection efficiency. Each of analyzed models are explained in detail below.

For the A3DR model, the Scaled Dot-Product Attention and Multi-Head Attention modules were used to build the model's attention mechanism. The model can concentrate on salient characteristics in video data thanks to its architecture, which is based on a 3D-ResNet-152 backbone and ensures that the model catches long-range temporal data dependencies.

In the T3DR model's architecture, the ResNet101 backbone is coupled with a Transformer encoder. The primary function of the encoder, which is constructed using a

Transformer-Encoder layer and a Transformer-Encoder module, is to capture long-range dependencies within video sequences. This architecture ensures effective feature encoding, which is pivotal for anomaly detection.

Finally, our ensemble model is a collection of five separate 3D-ResNet-152 Transformer models. Every model in the ensemble sees the data slightly differently, capturing various aspects of it. When combined, these models work together, resulting in better performance than any single model working alone.

6.4.1 Attention 3D-ResNet-152: Enhancing Feature Relevance and Discrimination

Combining attention processes with the 3D-ResNet-152 architecture improves feature relevance and discrimination in the A3DR model MIL [25]. Thus, merging these two techniques enables the model to choose concentrate on crucial elements while omitting less important data, improving anomaly detection performance.

DL models' attention processes let them give certain areas of the input data different weights. The attention mechanism assigns distinct weights to each feature channel in the 3D-ResNet-152 architecture [135], in the context of the A3DR. These feature channels describe particular traits or data patterns.

To mathematically represent this weighting process, consider equation 6.8 below:

$$w_i = \frac{\exp(f(x_i))}{\sum_{j=1}^{N} \exp(f(x_j))} \tag{6.8}$$

In this equation, w_i is the weight assigned to the i-th feature channel. The function $f(x_i)$ represents an activation function applied to the i-th channel, and N denotes the total number of feature channels in the model. This formula calculates the normalized weight for each feature channel based on its activation. The exponential function, exp, ensures that the weights are non-negative and emphasizes channels with higher activations.

The attention mechanism focuses the model on the most pertinent and discriminative qualities for anomaly detection by giving greater weights to more relevant feature channels and lower weights to less significant ones. Through this method, the model is able to give priority to the identification of abnormal patterns, which are essential for telling anomalous patterns apart from regular ones.

The benefit of employing attention strategies in this situation is that they enable the ability to focus on pertinent information, thereby significantly lessening the influence of unimportant or less discriminative qualities. As a result, the model gets more adept at catching and comprehending the minute variations between normal and abnormal patterns, which in turn improves the model's capacity to identify anomalies in video surveillance data.

6.4.2 Transformer 3D-ResNet-152: Capturing Long-Range Dependencies

The T3DR model combines the Transformer and the 3D-ResNet-152, two potent DL architectures. In order to identify odd or abnormal events in a video sequence, it is made to handle video data for the task of anomaly detection.

Recognising complicated patterns and anomalies in movies requires a grasp of the context and interactions between frames that are spaced widely apart in time. Conventional 2D CNNs have difficulty properly capturing long-range relationships. Modelling long-range relationships in text data has shown to be a strong suit of the Transformer architecture, which was initially created for natural language processing. This is so that each element in a sequence of Transformers can attend to other components regardless of their position by using mechanisms called self-attention mechanisms. Transformers are therefore a good choice for activities that call for the documentation of contextual relationships.

The self-attention mechanism of the Transformer can be mathematically expressed as follows in equation 6.9:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{6.9}$$

where Q, K, and V are the query, key, and value matrices derived from the input data, and d_k represents the dimension of the keys.

A 3D version of the well-known ResNet-152 CNN architecture, the 3D-ResNet-152 was created exclusively for video data. It can quickly process video frames while utilising 3D convolutions to capture spatial details. It cannot, however, model long-range temporal connections.

The 3D-ResNet-152 backbone and the Transformer's self-attention mechanism have been integrated by the researchers to overcome this constraint. Combining 3D-ResNet-152's spatial feature extraction skills with the Transformer's long-range contextual modelling capabilities enables the model to take advantage of the best of both worlds.

The resulting model, named T3DR, can successfully capture temporal dependencies and contextual information in video sequences by applying the Transformer architecture to the 3D-ResNet-152 backbone. The self-attention mechanism allows the model to concentrate on the most important frames at each time step and take long-term into account the relationships between frames.

Transformers' self-attention mechanism enables the model to dynamically determine the relative relevance of different frames in the video stream. This is important in the context of anomaly identification because aberrant occurrences may appear as departures from the usual patterns, which can be seen in the video over extended periods of time. The T3DR model is more able to recognise and detect abnormal occurrences than conven-

tional 3D CNNs because it focuses on the most important frames and captures long-range dependencies.

6.4.3 Ensemble of 3D-ResNet-152 Transformer Models: Leveraging Collective Knowledge

The accuracy and performance of anomaly detection can be significantly enhanced by applying ensemble learning [136], a powerful technique in the context of video surveillance anomaly detection, to models like the 3D-ResNet-152. By combining several individual models, often referred to as base learners or weak learners, through ensemble learning, it is created a more reliable and accurate ensemble model.

As explained before, the 3D-ResNet-152 is a DL architecture specifically designed for video analysis.

In this ensemble, each individual 3D-ResNet-152 model is trained on a specific subset of the training data or with a different initialization, allowing the ensemble to capture a diverse range of perspectives on irregularities identified in video surveillance data.

Our ensemble model consists of five independently trained 3D-ResNet-152 Transformer models. Each model is tailored to identify different aspects of video data, thereby enhancing the diversity in learning and detection capabilities. The final output is derived by averaging the predictions from all models, leveraging their collective intelligence for improved accuracy and robustness. This ensemble approach not only capitalizes on the varied insights of each model but also allows for dynamic adaptation to evolving data patterns, ensuring a more reliable and effective anomaly detection system overall.

The projections from each model in the ensemble are combined to yield the final conclusion, typically using voting or averaging methods. For binary classification tasks, the ensemble model might classify an instance as an anomaly if the majority of the base learners do so.

The ensemble learning process can be formalized using the following equations. Let's denote the ensemble as $\mathcal{E} = \{M_1, M_2, \ldots, M_k\}$, where M_i represents the i-th 3D-ResNet-152 model in the ensemble. The prediction of a base learner M_i for a given video sequence x is denoted as $p_i(x)$, whereas the final prediction of the ensemble model for video sequence x, denoted as $p_{\text{ensemble}}(x)$, is obtained using different aggregation methods, such as:

- **Voting**: in this approach, the final prediction is determined by a majority vote among the individual models, as shown in equation 6.10:

$$p_{\text{ensemble}}(x) = \text{argmax}_c \sum_{i=1}^{k} \mathbb{I}(p_i(x) = c), \tag{6.10}$$

  where c represents the class label, and $\mathbb{I}(.)$ is the indicator function.

- **Averaging**: in this approach, the final prediction is obtained by averaging the predictions of the individual models, as calculated in equation 6.11:

$$p_{\text{ensemble}}(x) = \frac{1}{k} \sum_{i=1}^{k} p_i(x). \tag{6.11}$$

The E3DRT models combines the outputs of different 3D-ResNet-152 Transformer models to increase anomaly detection performance. Trained on the UCF-Crime dataset, each model contributes to the ensemble by reducing reliance on any single model's biases or defects, thereby enhancing the overall stability and accuracy of the detection system.

The diversity of the models in the ensemble also plays a critical role in its success. Each model may excel in certain conditions or specialize in capturing specific anomaly characteristics. By merging the benefits and perspectives of multiple models, the ensemble provides a more comprehensive understanding of the video data, improving anomaly detection accuracy.

Empirical results and thorough evaluation have demonstrated that the E3DRT models consistently outperforms individual models, such as Attention 3D-ResNet-152 and T3DR, in terms of accuracy, robustness, and generalization capabilities.

By fusing ensemble learning with 3D-ResNet-152 Transformer models, we can benefit from the advantages of improved detection accuracy and robustness in anomaly detection for video surveillance applications.

6.5 Training Procedure and Model Evaluation

DL models resemble sculptured works of art. They start out as unformed, uncooked things that gradually take shape. Training is the process that creates these models, giving them meaning and functionality. It is during this vital stage that models develop from simple algorithms into strong instruments that can recognise patterns, identify anomalies, and more. We go further into this topic in this chapter.

It takes more than just feeding a model data to train it. It is important to comprehend the data, choose the best architecture for the job, set the appropriate parameters, and then adjust those parameters in response to input. This feedback is frequently described as a *loss value*. A model performs better the lower the loss. Besides, in this chapter, we will go into great depth on the quest to lessen this loss.

As stated before, the A3DR Model, the T3DR Model, and the Ensemble 3D-ResNet-152 Transformer Model are the three models we have selected for our study. Each model has a distinct methodology and offers various benefits for detecting video anomalies.

The UCF-Crime dataset [5], which is designed primarily for the identification of video anomalies, is used in the training of the three analyzed models. This dataset has been described in detail in section 3.2.

For the three approaches, a few critical parameters guide the training process:

- **Batch Size: 64**, chosen for striking a balance between gradient accuracy and computational performance. Smaller batches might not adequately reflect the variability of the dataset, while larger ones might be computationally expensive.
- **Feature Dimension: 2048**, the higher dimensionality allows the model to represent intricate patterns in the data, leading to better feature extraction and, subsequently, improved anomaly detection.
- **Number of Epochs: 150**, determined through validation performance, ensuring neither underfitting nor overfitting. This epoch count offers an optimal balance between training time and model performance.
- **Optimizer: Adadelta** [126], with an epsilon value of 1×10^{-8}. Adadelta is chosen for its adaptability to parameter updates, which can lead to faster convergence without manual learning rate tuning.

Each model undergoes training on the dataset for 150 epochs. Performance is optionally assessed using validation data. During training, the model's weights are iteratively updated using backpropagation in combination with the Adadelta optimization strategy. Monitoring involves tracking the 'loss' values throughout the training and validation phases. Callback functions are set up during training, providing real-time insights and control over the training process, such as early stopping or model checkpoints.

Figure 6.3 showcases the training loss evolution for the A3DR model. From the graph, it's evident that as the epochs progress, the loss decreases, suggesting the model is learning and improving its anomaly detection capabilities.

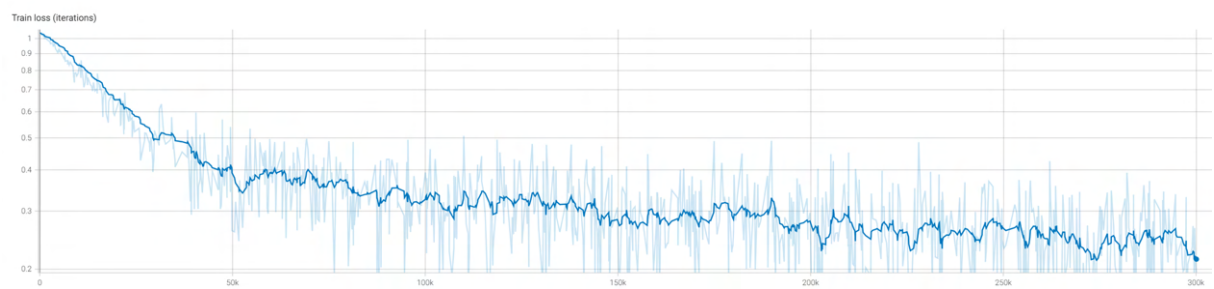


Figure 6.3: Training loss curve obtained with the A3DR.

Through this rigorous training process, the A3DR model, when fine-tuned on the UCF-Crime dataset, can effectively detect anomalies in video surveillance.

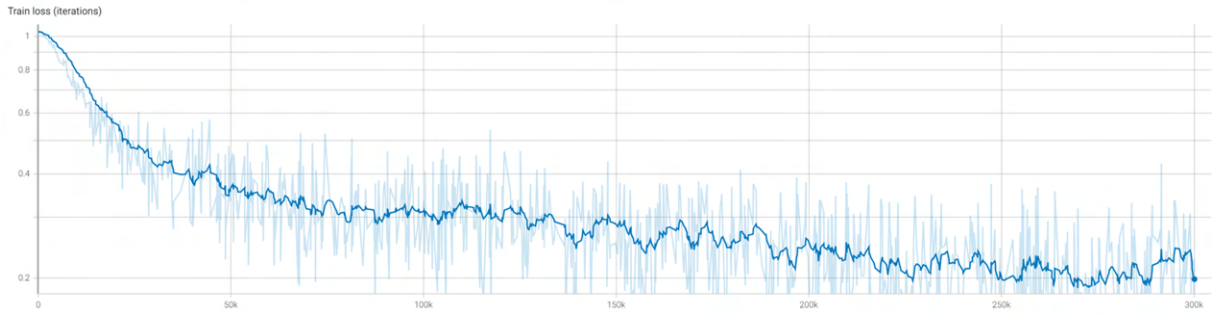


Figure 6.4: Training loss curve obtained with the T3DR.

Figure 6.4 demonstrates the decrease in training loss across epochs, indicating successful learning and anomaly detection capability development in the model.

Each of the five 3D-ResNet-152 Transformer models in the ensemble is trained on its own. They learn from the data via backpropagation and make adjustments with the help of the Adadelta optimizer.

In figure 6.6, we can see how each model improved during training. This is shown by the decline in training loss.

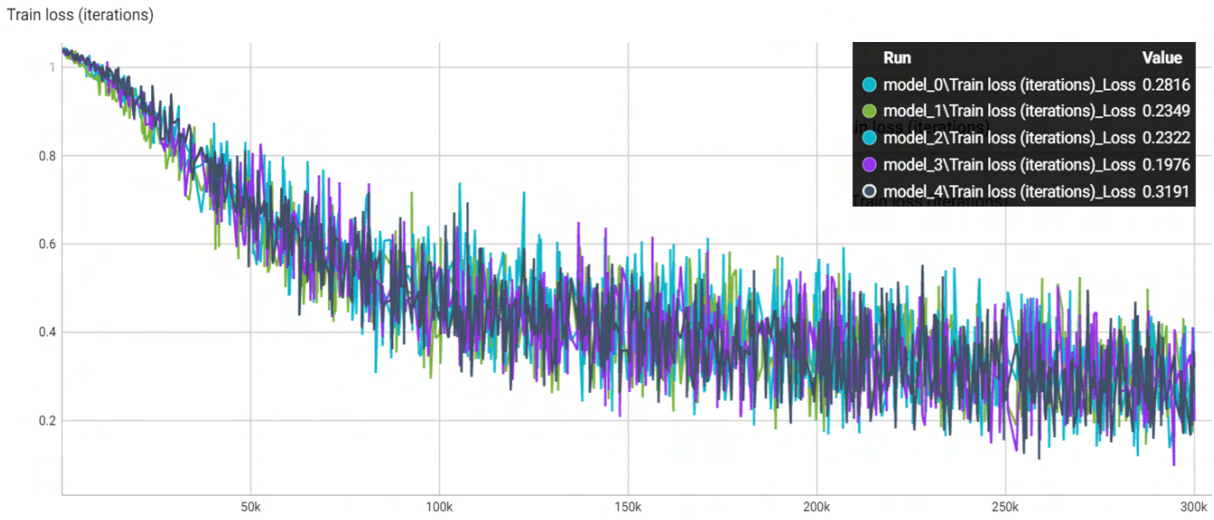


Figure 6.5: Training loss curve obtained with the E3DRT.

Looking at the training and validation loss curves, the final training and validation losses for the models are as shown in Table 6.1.

From the table, we observed that Model 3 not only achieved the lowest training loss but also maintained competitive validation loss, suggesting robust performance and good generalization compared to the other models. While Model 3 appears to be the most efficient at minimizing training loss, Model 4, despite having the highest training loss, shows the lowest validation loss indicating its potential to better generalize on unseen data. This disparity highlights the value of considering both training and validation performance in evaluating models. The variation in performance metrics among the models is beneficial

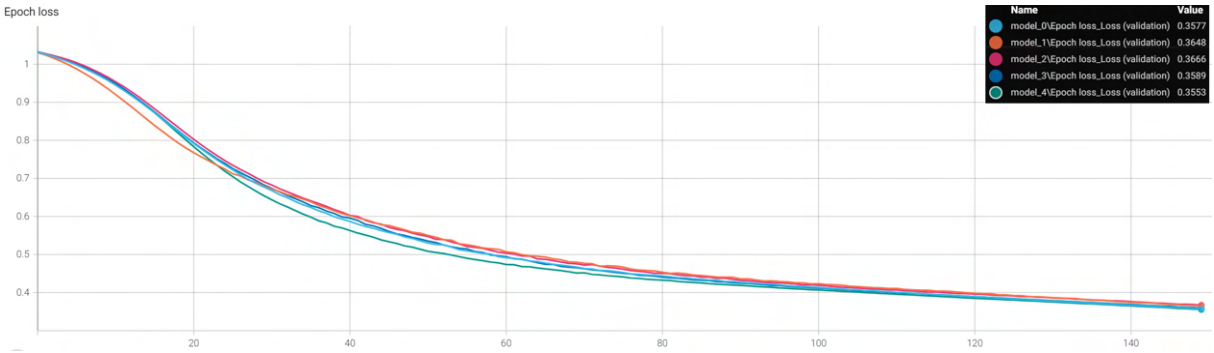


Figure 6.6: Epoch Loss(validation) obtained with the E3DRT.

Table 6.1: Final training and validation losses for the models in the E3DRT.

Model	Training Loss	Validation Loss
Model 0	0.2816	0.3577
Model 1	0.2349	0.3648
Model 2	0.2322	0.3666
Model 3	**0.1976**	0.3589
Model 4	0.3191	0.3553

in an ensemble setup, as each model brings a different perspective, enhancing the ensemble's overall robustness and effectiveness in detecting anomalies in video surveillance data.

6.6 Real-time Anomaly Detection Architecture

The rapid proliferation of video surveillance systems has underscored the indispensable need for efficient and timely anomaly detection. While conventional systems have presented a series of solutions, their inherent limitations in dealing with live-streams have often led to inefficiencies. Addressing these gaps, our proposed real-time anomaly detection architecture (a lighter version based on the approaches explained in previous sections) offers a potent blend of temporal segmentation and advanced DL techniques. Before diving into the intricacies of our methodology, highlighted in figure 6.7, it is crucial to understand its genesis and overarching structure. The final architecture is discussed in detail as follows:

Our framework commences by acquiring a video stream as its primary input. To ensure computational efficiency without compromising significantly on information retention, each frame from this video undergoes resizing to a more manageable resolution of 112×112 pixels. Such a reduced resolution paves the way for diminished computational overhead and memory usage, vital for real-time processing.

The essence of videos lies in their temporal flow. Recognizing this, the resized video undergoes temporal segmentation, wherein it's subdivided into batches comprising 16

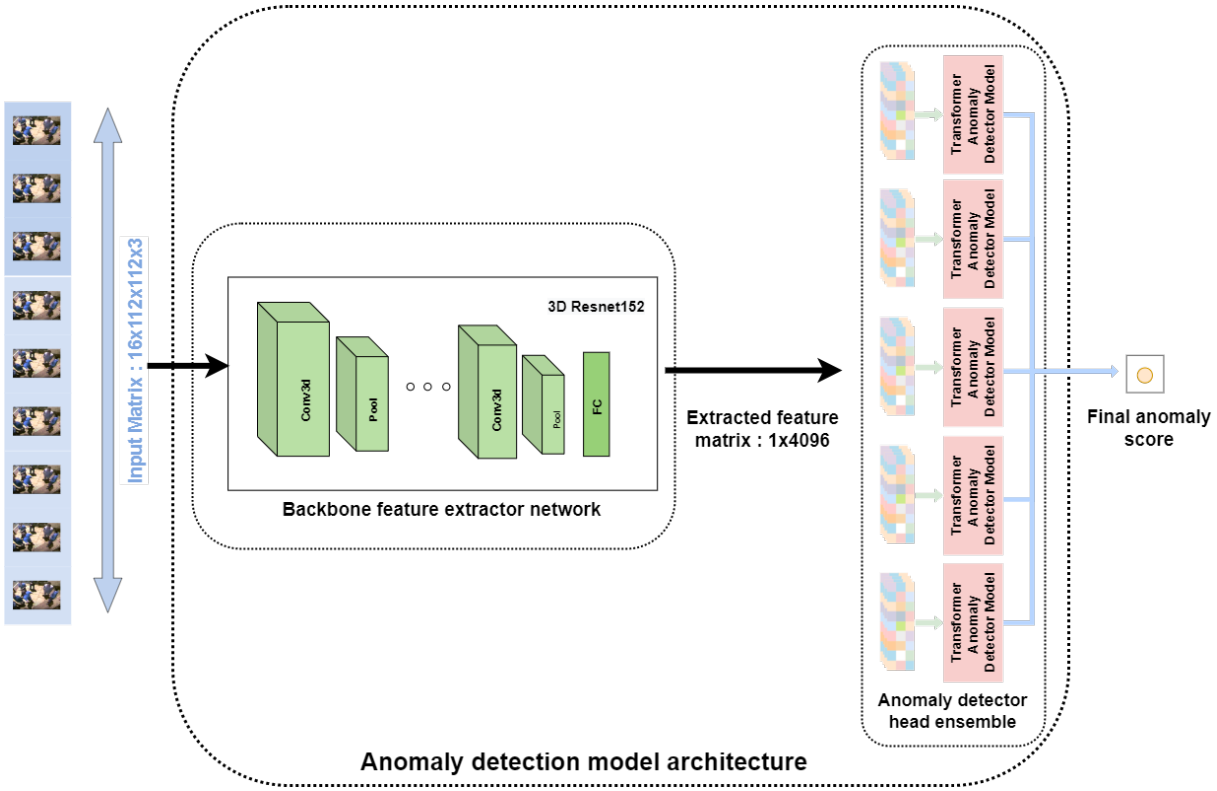


Figure 6.7: Proposal of real-time anomaly detection architecture.

consecutive frames, ensuring the retention of the temporal lineage. Opting for a batch size of 16 frames, as opposed to larger sizes like 32 or 64, balances computational efficiency with temporal information richness. This ensures that the model remains responsive in a live-feed scenario while still capturing the short-term temporal dynamics effectively.

Subsequently, each of these temporally segmented batches is structured as a 4-dimensional matrix with dimensions $16 \times 112 \times 112 \times 3$, where:

- 16: denotes the number of frames per batch, symbolizing the temporal context.
- 112×112: means the height and width of every resized frame.
- 3: represents the standard RGB color channels, ensuring a comprehensive color-based representation for each pixel.

At the heart of the architecture is the 3D-ResNet-152, tailored for processing spatio-temporal data. This network ingests the $16 \times 112 \times 112 \times 3$ matrix and conducts convolutions and pooling operations across both the spatial and temporal realms. Through these intricate operations, the network discerns and teases out salient spatio-temporal features from the video segments.

Post the operations of the 3D-ResNet-152, we are presented with a 1-dimensional feature vector of dimensions 1×4096. This compact vector, despite its reduction in size, encapsulates critical spatio-temporal patterns and nuances of the original temporal batch.

Harnessing the power of Transformer-based architectures, renowned for their prowess with sequential data, the extracted 1×4096 feature vector is probed by an ensemble of 5 such models. This ensemble approach amplifies the robustness of the detection process by pooling insights from multiple models.

For the given feature vector, each model in the ensemble provides an anomaly score. These scores are not handled equally. Instead, a weighted amalgamation technique is used, which may provide models with better performance during training or validation phases more weights. Based on the ingested spatio-temporal relations and representations, this weighted anomaly score provides a measurable value of the temporal batch's anomaliness.

Additionally, our prototype includes a GUI, because we recognise that usability and interpretability are crucial for the adoption of such systems. This GUI helps users visualise and comprehend the underlying processes and outcomes in addition to giving them a physical interface to interact with the system. A detailed dissection of the GUI, along with a comprehensive analysis of our prototype's performance, is elucidated in section 6.7.

6.7 A GUI for the Real-time Anomaly Detection Prototype

This part covers the development of a desktop program (written in Python) including a GUI. The purpose of the program is to provide a user interface to test how each model works with anomaly detection tasks in real time. It works as follows:

Figure 6.8 illustrates the first window shown to the user. In this window, the user is presented with the following options:

- **Model Selection:** where the user can choose between three different models for performing the anomaly detection. The available options are:
  - Attention 3D-ResNet-152 Model
  - Transformer 3D-ResNet-152 Model
  - Ensemble 3D-ResNet-152 Transformer Model

  The user can select the desired model by clicking on the corresponding button.
- **Camera Selection:** the user can choose a specific surveillance camera input for the anomaly detection. This allows the user to test the program with different camera feeds. The available cameras are listed in a drop down menu, and the user can select the desired camera from the list.
- **Record:** the user has the option to record the anomaly detection session. By clicking on the **'Record'** button, the program will start recording the video feed along with the anomaly detection results. The recorded video will be saved in `MP4` format.

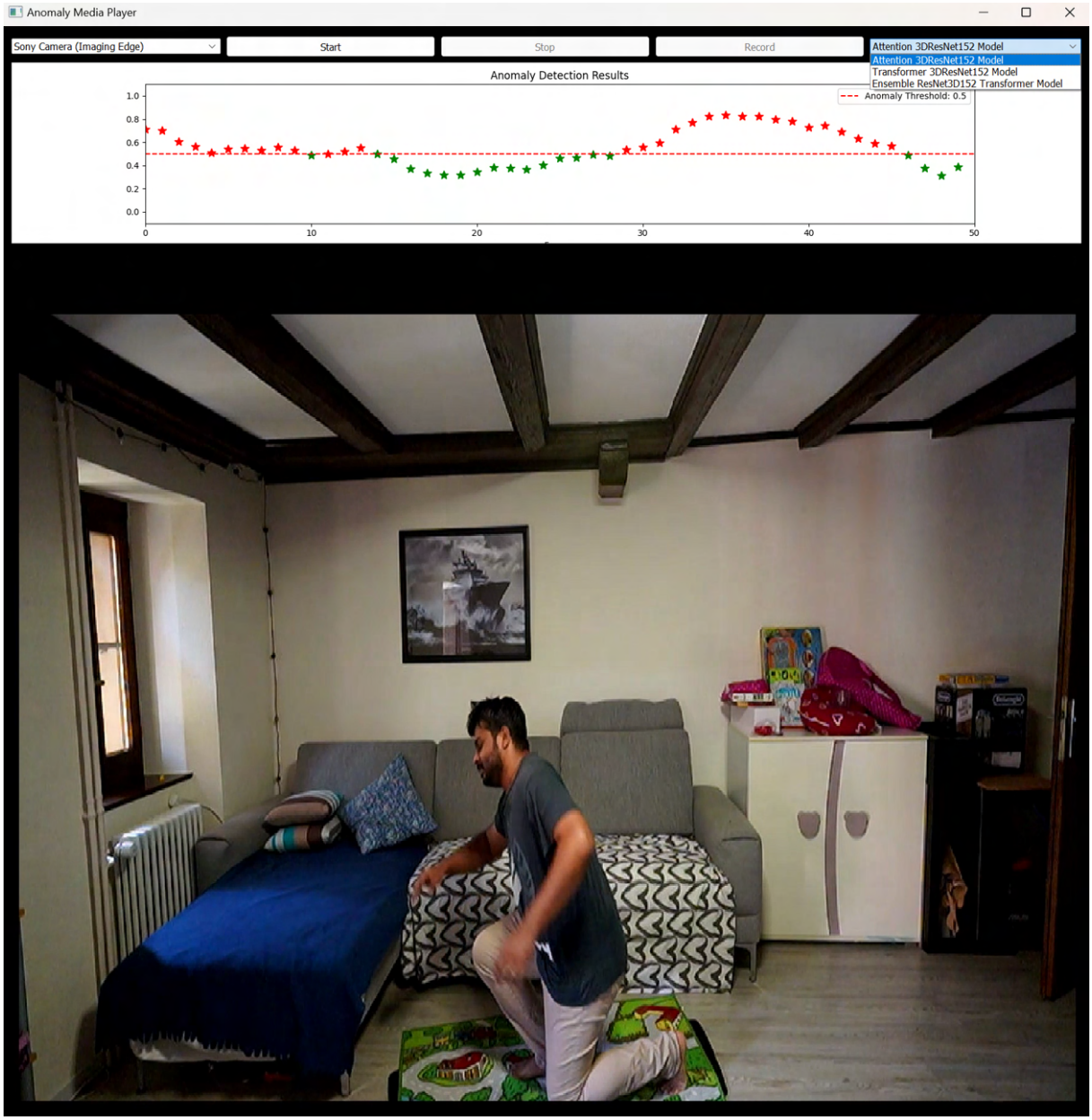


Figure 6.8: Appearance of the real-time anomaly detection proposal GUI.

Once the model has been selected and a specific surveillance camera input is chosen, the user can press the **'Start'** button to initiate the anomaly detection process. As a result, two additional windows will appear at the bottom of the screen:

- **Classification Window:** this window displays the classification results of the anomaly detection algorithm in real time. Each star represents a single frame, and the color of the star indicates the anomaly score of that frame. In normal scenarios, the stars will appear in green color, indicating a low anomaly score. However, in anomaly situations, the stars will appear in red color, indicating a high anomaly score. The classification window provides a visual representation of the anomaly detection results.

- **Video Feed:** this window displays the real-time video feed from the selected surveillance camera. The video feed is used as input for the anomaly detection algorithm, and the user can observe the detected anomalies in real time.

The availability of three different models for selection allows the user to choose the most appropriate model based on their specific requirements. In our findings, each model performs differently in different scenarios, so the users can select the model that best suits their needs.

By providing a GUI with real-time anomaly detection capabilities, our prototype aims to facilitate the testing and evaluation of different anomaly detection models in practical surveillance scenarios. The user can interact with the program, observe the anomaly detection results, and make informed decisions based on the performance of each model.

6.8 Conclusions

We have analyzed the characteristics of A3DR model, T3DR model, and the E3DRT model on the UCF-Crime dataset by examining the features, benefits, and potential downsides of each model. Moreover, the performance of each model is analyzed in chapter 6.8.

A 3D-ResNet-152 backbone and an attention mechanism are combined in the A3DR model. The attention mechanism is implemented via the Scaled Dot-Product Attention [24] and the Multi-Head Attention modules [24], which helps the model concentrate on crucial details and identify long-term temporal correlations in the video data. In the realm of VAD, the ability to recognise important events and trends across time is particularly useful. The model might not function as effectively when the abnormal patterns are obscure or very changeable.

The T3DR model combines the Transformer encoder with 3D-ResNet-152 backbone. While capturing long-range dependencies, the Transformer encoder effectively executes feature encoding for anomaly detection. This model benefits from the Transformer's ability to analyse complete sequences, understand context, and identify correlations in the data, which is crucial for identifying sporadic or intricate anomalies. The complexity of the Transformer model, however, may necessitate a longer training period and more computational work.

The E3DRT model trains multiple 3D-ResNet-152 Transformer models using an ensemble method. It is possible to combine these models to make use of their combined advantages, perhaps improving performance. This model is more robust against different kinds of anomalies than other ones because of the robustness of its views. Because it can be time- and resource-intensive to train an ensemble of numerous models, the cost of computation is a trade-off.

When considering the UCF-Crime dataset, which consists of a variety of real-world surveillance footage with various types of abnormalities, the E3DRT Model provides the best results. This approach combines the strengths of ensemble learning, which can produce more dependable performance when dealing with a wide range of anomalous events, with the advantages of the Transformer's capacity for contextual and sequential knowledge.

Furthermore, real-time VAD has reached a turning point with the introduction of our suggested architecture. The combination of temporal segmentation and modern DL paradigms is an example of a forward-thinking strategy that has been painstakingly designed to capture the latent complexities of live video feeds.

The architecture's various components, ranging from the initial video intake to the extraction of subtle spatio-temporal signals and subsequent ensemble-based anomaly detection, all showcase a commitment to robustness and real-time responsiveness. By integrating an ensemble of Transformer-based models, an enhanced layer of redundancy and resilience is introduced, significantly reducing the likelihood of missing anomalies.

Results for these techniques are shown in chapter 6.8, along with comprehensive performance metrics.

6.9 Experimental Setup

In the Experimental Setup of our study, we designed a framework to evaluate the efficacy of our proposed anomaly detection algorithm. Our experiments leveraged a high-performance computing environment, consisting of a robust system equipped with an AMD Ryzen 9 5900X 12-Core Processor operating at 3.70 GHz, and enhanced by an NVIDIA RTX 3080 GPU. This configuration was essential for handling the intensive computational demands of our sophisticated deep learning models.

Our analysis incorporated two comprehensive datasets, the ShanghaiTech dataset and the UCF-Crime dataset (discussed in detail in section 3). These datasets are recognized for their diverse range of real-world surveillance footage and challenging anomaly detection scenarios, providing a realistic testing ground for our models. Moreover, GBA dataset is also used to test the proposal in the wild.

To ensure consistency and fairness in our comparisons, we replicated the training conditions of benchmark methods. This included maintaining uniformity in epochs, batch sizes, and learning rates. The chosen evaluation metrics, primarily focusing on frame-level AUC performance, were selected to accurately reflect the accuracy and reliability of anomaly detection in our models.

This comprehensive setup provided a robust platform for assessing our algorithm's performance and allowed for a direct comparison with existing state of the art methods. This approach was instrumental in establishing a thorough understanding of our proposed models' capabilities in real-world surveillance scenarios.

6.10 Performance Evaluation

The experimental results of our proposed anomaly detection algorithm on ShanghaiTech and UCF-Crime datasets, presented in table 6.2, demonstrate its effectiveness and outperformance compared to previous methods.

For ShanghaiTech dataset, our proposed method achieves superior results when compared to unsupervised learning methods [4] and weakly supervised approaches Sultani *et al.* [5]. Particularly, using the Attention 3D-ResNet-152 architecture, our proposed method achieves an accuracy of 91.46%. This highlights its ability to accurately detect anomalies in ShanghaiTech dataset.

Analyzing the results for the UCF-Crime dataset, the proposed method outperforms all previous unsupervised learning approaches. Remarkably, using the Attention 3D-ResNet-152, Transformer 3D-ResNet-152, and Ensemble 3D-ResNet-152 Transformer architectures, it achieves accuracies of 88.28%, 82.38%, 83.53%, and 88.28%, respectively, as shown in table 6.2. These results surpass the performance of current state of the art MIL-based methods.

It is worth noting that the proposal in the second contribution based on the 3D-ResNet-152 with attention mechanism obtains better results for UCF-Crime that the Attention 3D-ResNet-12 in the final proposal because we used multi-head channel-based attention in the second contribution and used channel-based attention in the final proposal.

Figure 6.9 presents the ROC comparison of the three proposed approaches for VAD, providing a concise visual summary of their performance for UCF-Crime dataset.

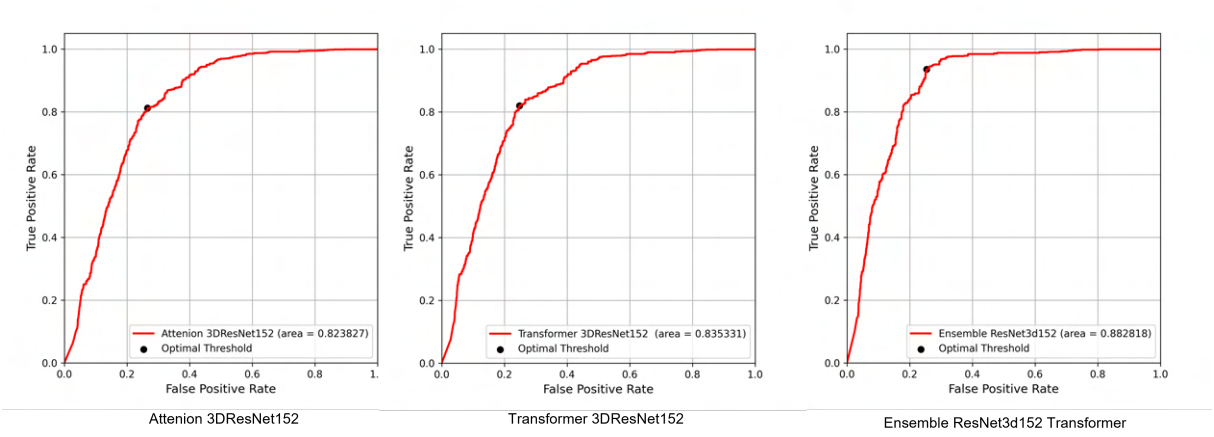


Figure 6.9: ROC of the three proposed models for anomaly detection in UCF-Crime dataset.

The significant improvement in performance on both datasets validates the effectiveness of our proposed algorithm. By leveraging advanced DL architectures, such as Attention 3D-ResNet-152 and Transformer 3D-ResNet-152, our model demonstrates superior anomaly detection capabilities. These results highlight the potential of our proposed al-

Table 6.2: Comparison of frame-level AUC performance with other state of the art unsupervised and weakly supervised methods on ShanghaiTech (column 4) and UCF-Crime (column 5) datasets.

Supervision	Method	Feature	AUC (%)	
			ShanghaiTech	UCF-Crime
Unsup.	Stacked-RNN [110]	-	68.00	-
	FFP [4]	-	73.4	-
	Mem-AE [113]	-	71.20	-
	VEC [119]	-	74.80	-
	Lu *et al.* [109]	C3D-RGB	-	65.51
	GODS [15]	I3D-RGB	-	70.46
Weakly Sup.	GCN [117]	C3D-RGB	76.44	81.08
		TSN-Flow	84.13	78.08
		TSN-RGB	84.44	82.12
	Zhang *et al.* [116]	I3D-RGB	82.50	-
		C3D-RGB	-	78.66
	Sultani *et al.* [5]	C3D	-	75.41
	Motion Aware [9]	PWC Flow	-	79.00
	1^{st} **Contribution [23]**	T-C3D-$\log_2$	-	80.36
	2^{nd} **Contribution [25]**	3D-ResNet-152 Att.	-	87.65
		I3D-Att.	91.21	83.69
	Final Proposal	A3DR	**91.46**	82.38
		T3DR	-	83.53
		E3DRT	-	**88.28**

gorithm in enhancing surveillance and security systems, enabling accurate and efficient anomaly detection in real-world scenarios.

6.11 Qualitative Analysis

This section highlights the principal outcomes derived from our experimental evaluations. We employed a complex anomaly detection model for our tests that proved successful even in scenarios with subtle abnormalities that were barely visible. Despite the challenges associated with identifying anomalous events in real-world applications due to their infrequent occurrence, complexity, and unpredictability, our proposed model demonstrated significant proficiency in detecting multiple anomalous events within a single video.

6.11.1 Anomaly Detection Model Visualization

The illustrations provided in this section showcase the experimental results of the anomaly detection model as applied to various datasets. The primary dataset is the UCF-Crime dataset [5], supplemented by additional data from GBA dataset [121], The Web Dataset [54], and the ShanghaiTech dataset [4].

Figures in section 6.11.2 provide insights into the model's performance on the UCF-Crime dataset. Each figure represents the model's output for a distinct crime category.

In contrast, Figure 6.25 presents results from GBA dataset, whereas Figure 6.24 from The Web Dataset, and Figure 6.26 showcases results from the ShanghaiTech dataset. These results illustrate the model's versatility and effectiveness when applied to diverse datasets.

Significantly, while testing the model on GBA, The Web, and ShanghaiTech datasets, we did not use any training videos specific to these datasets. Instead, we utilized the training features derived from the UCF-Crime dataset to understand our model's generalization capabilities.

The visualized results are represented by colored circles and indication lines. Their interpretation is explained as follows:

- **Red circles:** mark the onset of an anomaly.
- **Green circles:** indicate normal events.
- **Yellow circles:** highlight instances with an anomaly incorrectly detected.
- **The indication line, typically green, turns red** when the anomaly score surpasses 50% a designated threshold (represented by the blue dotted lines).

This visual representation reveals the model's ability to discern and denote anomalous activities, even within complex video sequences. Instances where the model failed to detect anomalies are also marked (yellow circle), providing a comprehensive perspective on the model's performance.

We began our study by examining all 13 classes from the UCF-Crime dataset [5] as well as the normal class. This primary analysis was later expanded to include samples from the ShanghaiTech, The Web Dataset [54] and GBA dataset [121]. The aim was to prove the model's applicability across various contexts since it does not rely on any specific type of abnormal event.

6.11.2 Detailed Analysis on UCF-Crime Dataset

Qualitative visual results of 13 abnormal classes and one normal class from the UCF-Crime dataset [5] are presented and analyzed below.

6.11.2.1 Anomalous Classes in UCF-Crime

The anomalous classes are sorted alphabetically, whereas the normal one is shown at the end of this section.

1. **Abuse Class:** Figure 6.10 presents an example from the 'Abuse' class, which involved a scenario inside a police station room. In this example, our model detected the anomaly when an officer forcefully shut the door on a woman, represented by the red mark. The green mark signifies a normal event, where both the officer and woman exited the room without any abnormal activity. However, later in the video, when the officer re-entered the room and began assaulting the woman, our model successfully detected this anomalous event.

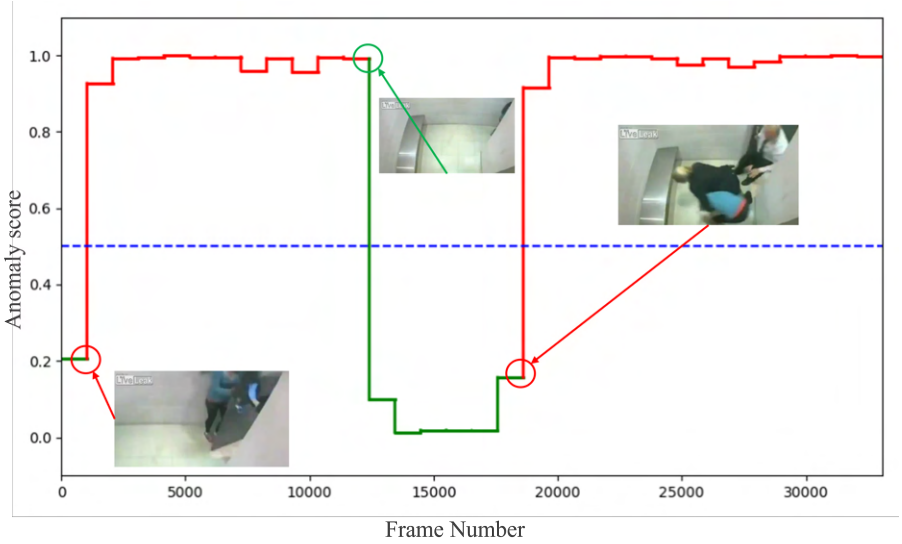


Figure 6.10: Qualitative visual result of 'Abuse' class in UCF-Crime dataset.

2. **Arrest Class:** In the second experiment from the 'Arrest' class, shown in Figure 6.11, a vehicle was being chased by the police. Initially, the vehicle forcefully tried to cross the lane, which was detected as an anomaly by our model and represented by a red mark. In the middle of the chase, the vehicle hit a police car head-on, which was also detected as an anomaly by our model and represented by a red mark. Finally, the video ended with no abnormal activity detected by our model and represented by a green mark.

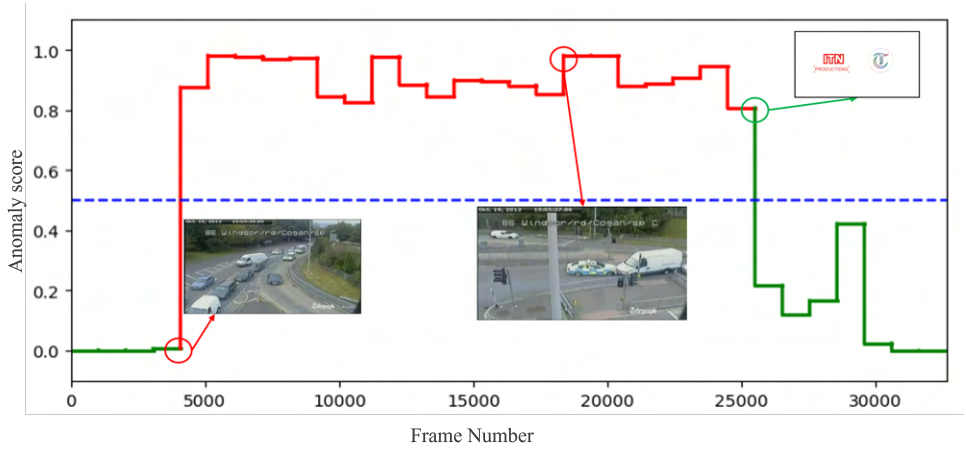


Figure 6.11: Qualitative visual result of 'Arrest' class in UCF-Crime dataset.

3. **Arson Class:** The third example, presented in Figure 6.12, corresponds the 'Arson' class. In this sequence, a person tried to throw a 'molotov-cocktail' to a store, which was detected as an anomaly by our model and represented by a red mark. After some time, a big fire light occurred, which was also detected as an anomaly by our model and represented by a red mark. In some phases of the video, the model incorrectly detected an anomaly, which was represented by a yellow mark.

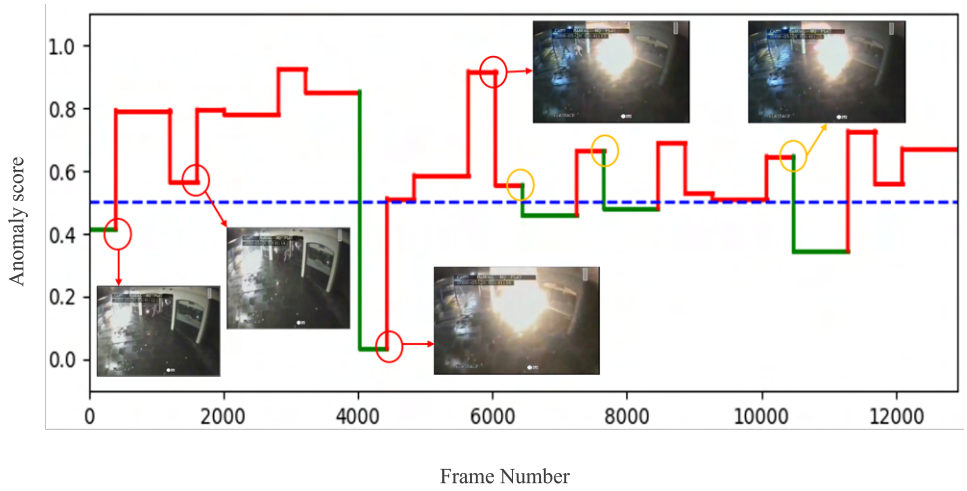


Figure 6.12: Qualitative visual result of 'Arson' class in UCF-Crime dataset.

4. **Assault Class:** In the fourth experiment from the 'Assault' class (Figure 6.13), a person followed a girl walking, which was detected as an anomaly by our model and represented by a red mark. After some time, the person hit the girl on the back of her head and she fell on the road, which was also detected as an anomaly by our model and represented by a red mark. In some phases of the video, the model incorrectly detected an anomaly, which was represented by a yellow mark. Finally, the video ended with no movement detected by our model while the girl was lying on the ground, which was represented by a green mark.

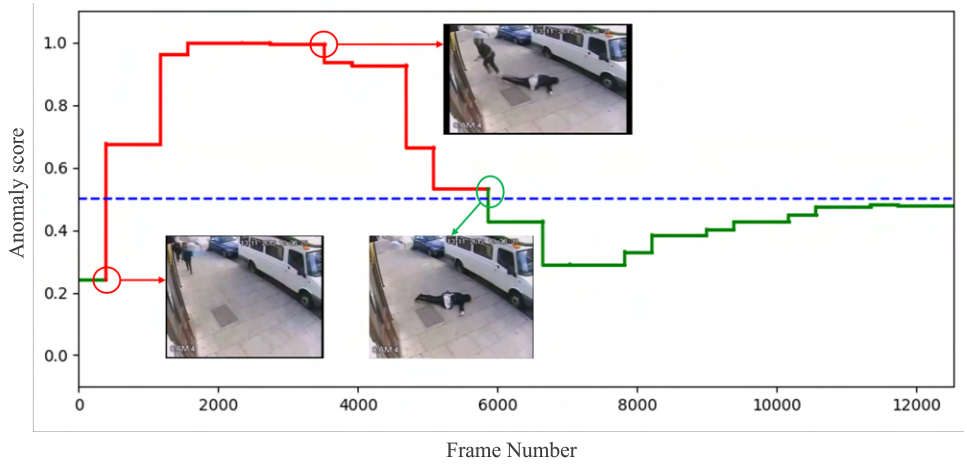


Figure 6.13: Qualitative visual result of 'Assault' class in UCF-Crime dataset.

5. **Burglary Class:** Figure 6.14 shows an example from the 'Burglary' class, in which our model successfully identified anomalies at multiple stages of the video. Firstly, during the initial phase, the model detects an anomaly represented by a red mark when a vehicle is about to collide with a gas station. This detection indicates a potential dangerous situation or irregular behavior. Next, as the video progresses, people exit a car and start running around, which our model recognizes as another anomaly, also marked in red. This detection suggests abnormal activity within the video footage, possibly indicating suspicious behavior or a potential threat.

   However, our video analysis model is not perfect and occasionally makes incorrect anomaly identifications. In the example, represented by a yellow mark, the model incorrectly detects an anomaly during a specific phase of the video. These instances highlight the areas where the model's performance may need further improvement. Finally, as the video reaches its conclusion, our model marks a portion of the footage with a green mark, indicating no movement or lack of any abnormal activity. This observation suggests that the video has returned to a normal state, with the vehicle remaining stationary.

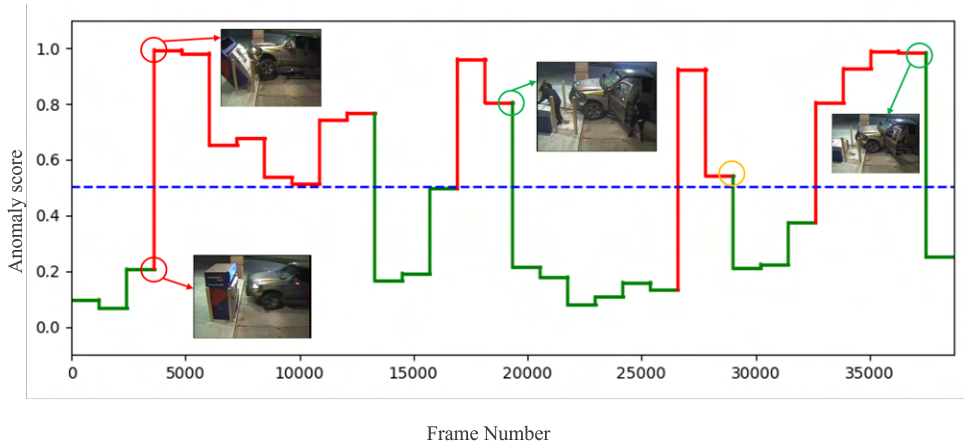


Figure 6.14: Qualitative visual result of 'Burglary' class in UCF-Crime dataset.

6. **Explosion Class:** The next example (Figure 6.15) corresponds to the 'Explosion' class. In this sequence, the video begins with a parked vehicle, indicated by a green mark representing no movement detected by our model. After a certain period of time, an explosion occurs in the background, which our model successfully detects as an anomaly, represented by a red mark. During some phases of the video, the model highlights instances where it inaccurately identified non-anomalies. These instances are depicted by yellow marks, indicating that the model incorrectly detected an anomaly in those frames. Finally, the video ends with a black screen, signifying that the Closed-Circuit Television (CCTV) system shuts down. The model correctly identifies this lack of movement as no anomaly, represented by a green mark.

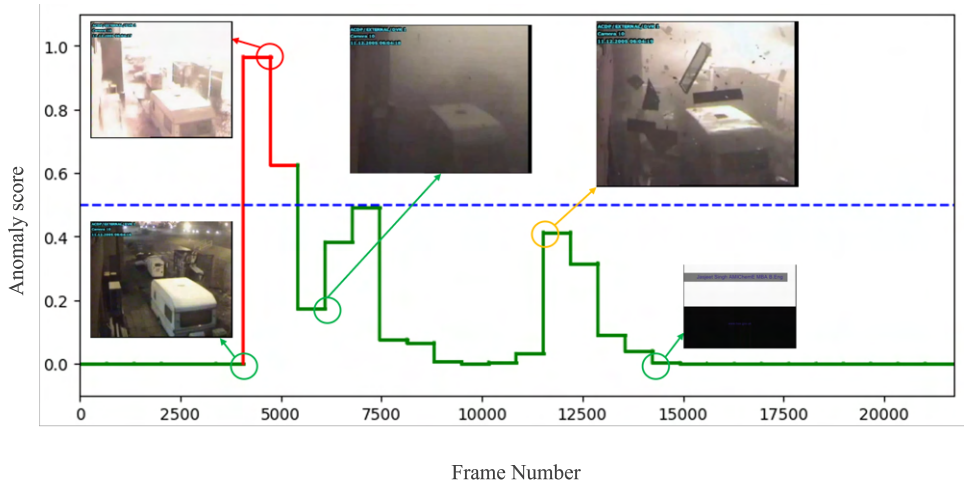


Figure 6.15: Qualitative visual result of 'Explosion' class in UCF-Crime dataset.

7. **Fighting Class:** In the seventh experiment from the 'Fighting' class, presented in Figure 6.16, our model identified abnormal movement between two individuals at the beginning of the video. This abnormal movement was detected and represented by a red mark, indicating a potential anomaly. As the video progressed, a fighting situation emerged. Our model successfully recognized this anomaly, depicted by another red mark. The system's anomaly detection capabilities were able to discern the unusual activity and flag it as a potential issue. Lastly, towards the end of the video, the scene transitioned into a no-movement state. Our model detected this lack of movement and marked it with a green indicator. This phase represented a normal state, contrasting with the preceding anomalies.

8. **Road Accident Class:** An example from the 'Road Accident' class in shown in Figure 6.17. In this example the video depicted a scene where people were walking, exhibiting normal movement. Our model successfully recognized this normal movement and represented it with a green mark, indicating a standard activity. As the video progressed, a critical event unfolded when a bus appeared on the scene, traveling at an excessive speed and ultimately crashing into a pole. This event was identified as an anomaly by our model, which promptly marked it with a red indica-

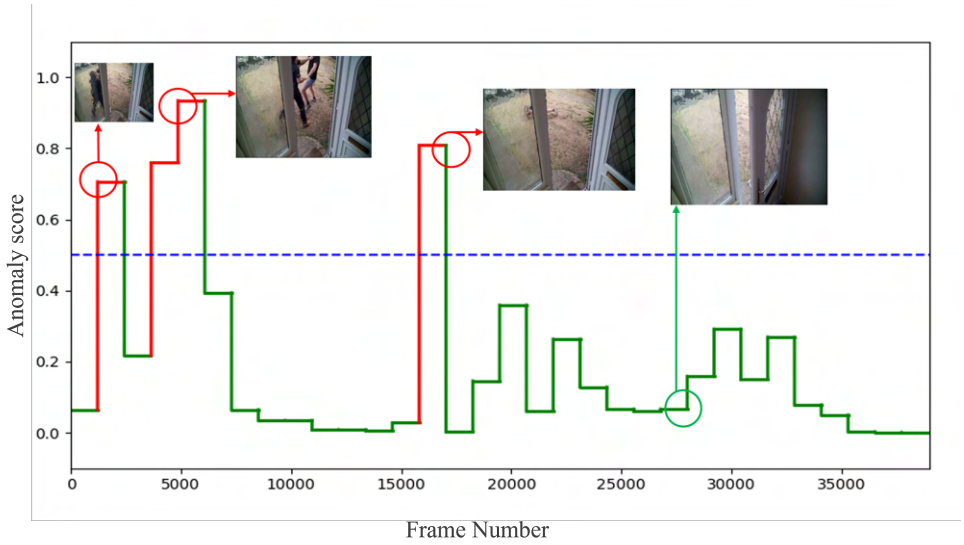


Figure 6.16: Qualitative visual result of 'Fighting' class in UCF-Crime dataset.

tor. The system's anomaly detection capabilities were successful in pinpointing the irregular occurrence, showcasing its ability to recognize and flag anomalies within video footage.

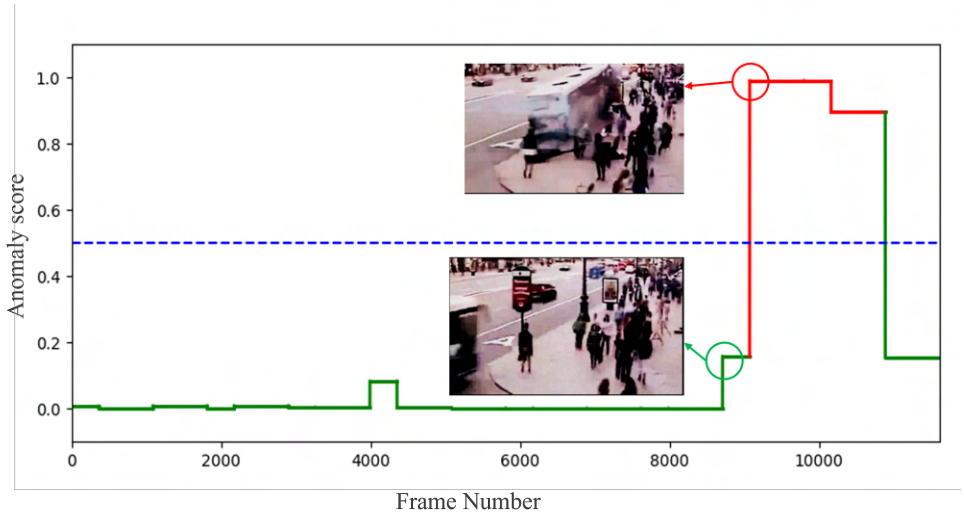


Figure 6.17: Qualitative visual result of Road 'Accident' class in UCF-Crime dataset.

9. **Robbery Class:** The ninth experiment (see Figure 6.18) correspond to the 'Robbery' class, in which our model detected the normal movement within the shop initially, represented by a green mark. It then identified the fighting between the shopkeeper and the robber as an anomaly, marked with a red indicator. Finally, the system recognized the absence of movement at the end of the video, denoting the conclusion of the robbery.

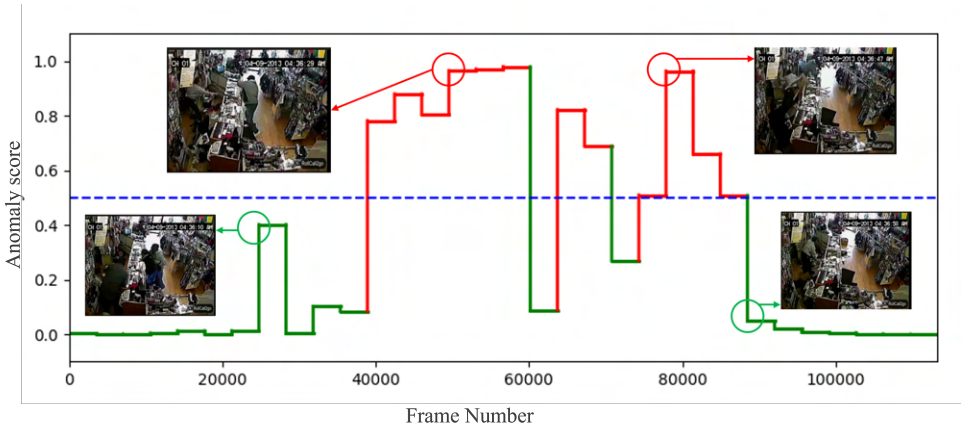


Figure 6.18: Qualitative visual result of 'Robbery' class in UCF-Crime dataset.

10. **Shooting Class**: Figure 6.19 presents an example from the 'Shooting' class. In this sequence, our model effectively detected the abnormal movement of the security person who was shot and sought cover, represented by red marks. As the video progressed, it reached a final phase where the shooting incident concluded. In this phase, there was no movement observed.

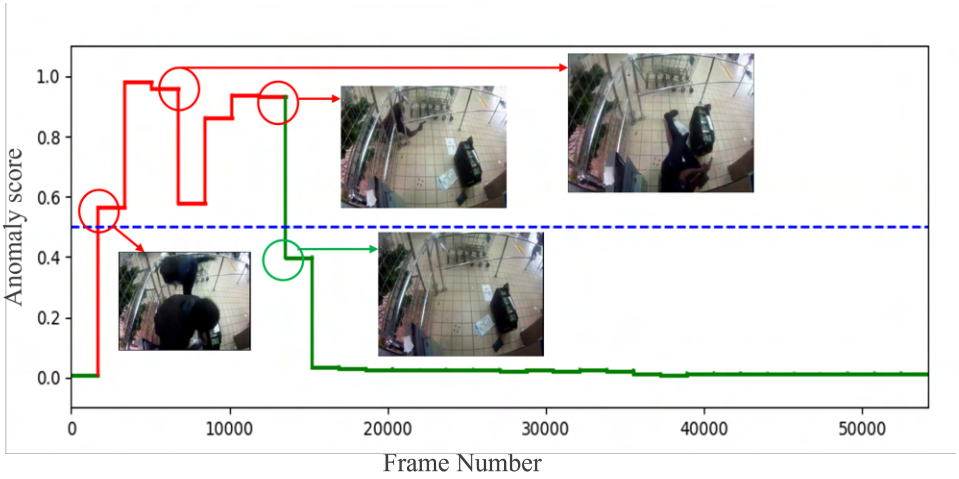


Figure 6.19: Qualitative visual result of 'Shooting' class in UCF-Crime dataset.

11. **Shoplifting Class**: A sequence from the 'Shoplifting' class is represented in Figure 6.20. As it can be seen, at the start of the video, there was an anticipation of abnormal movement by a female individual, indicating a potential shoplifting attempt. However, our model detected no abnormal movement during this phase and represented it with a green mark, suggesting that no anomaly was detected. As the video progressed, the female successfully carried out the theft and concealed the stolen product. Our model, in this case, erroneously failed to detect the anomaly, and no red mark was provided to indicate the abnormal activity. This represents an incorrect detection by the model, which is denoted by a yellow mark. In the final

phase, the female left the store without any further notable activity. The model identified the lack of anomaly in this phase and marked it with a yellow mark.

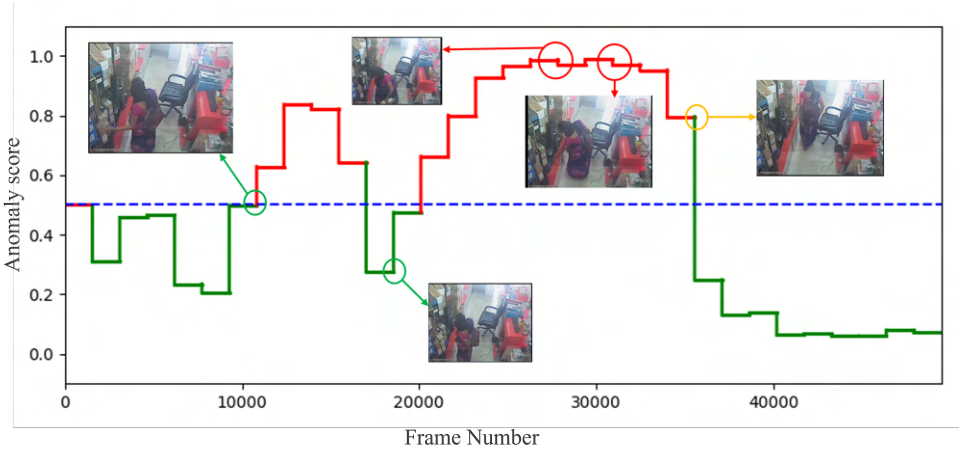


Figure 6.20: Qualitative visual result of 'Shoplifting' class in UCF-Crime dataset.

12. **Stealing Class:** In Figure 6.21, it can be observed an example belonging to the 'Stealing' class. At the beginning of the video, there was an indication of abnormal movement by an individual, suggesting an impending theft of a bike. Our model successfully detected this future abnormal movement and represented it with a red mark, highlighting the potential anomaly. As the video progressed, the person acted as if they were the owner of the bike and attempted to unlock it. Our model correctly identified that no anomaly was present during this phase, represented by multiple green marks indicating normal behavior. In the final phase, the person resorted to kicking the handlebar of the bike to break the neck-lock. However, our model inaccurately classified this action as a non-anomaly, resulting in an incorrect detection represented by a yellow mark. Ultimately, the person successfully stole the bike and left the scene while appearing to be the owner. Our model correctly identified the absence of any anomaly during this phase and marked it with a green indicator, signifying a normal detection.

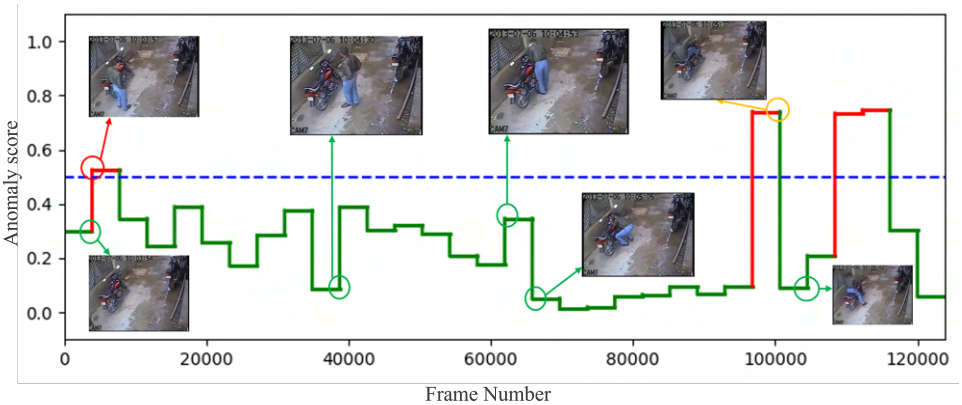


Figure 6.21: Qualitative visual result of 'Stealing' class in UCF-Crime dataset.

13. **Vandalism Class:** In the thirteenth experiment from the 'Vandalism' class, presented in Figure 6.22, the video starts by showing a person throwing an object at a property, and our model detected that behavior as an anomaly, represented by a red mark. While that person was walking, our model predicted it as normal behavior. Suddenly, that person again throws an object, and our model continuously predicted it as an anomaly.

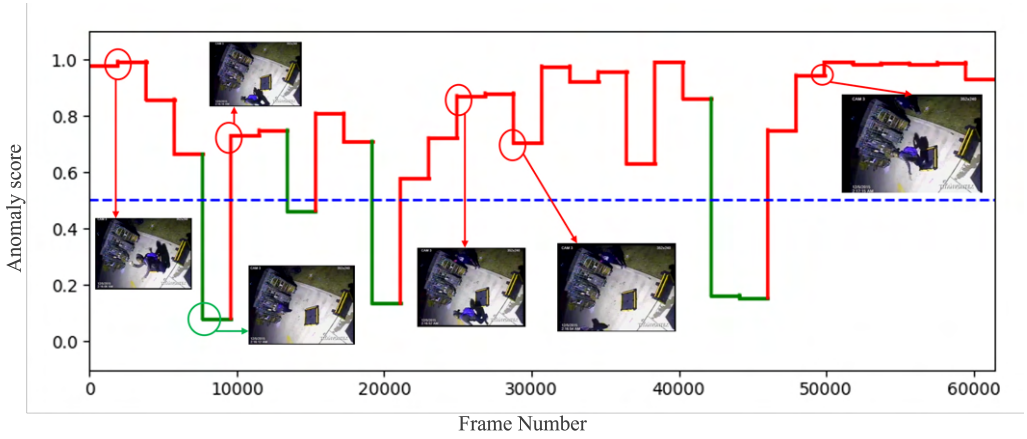


Figure 6.22: Qualitative visual result of 'Vandalism' class in UCF-Crime dataset.

6.11.2.2 Normal Class in UCF-Crime

Below, Figure 6.23 shows an example from the 'Normal' class. In that sequence, the video shows a mother and child walking towards home. Since there is no abnormality observed, our model predicted it as a normal video.

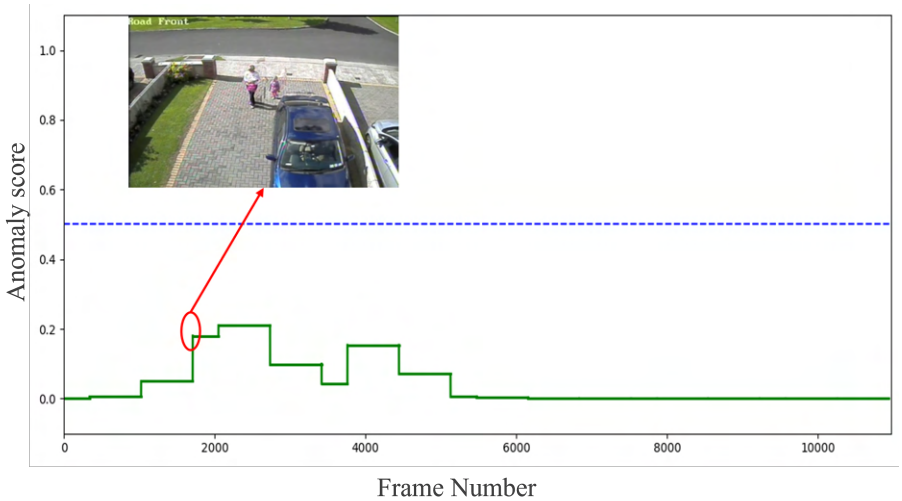


Figure 6.23: Qualitative visual result of a normal video in UCF-Crime dataset.

6.11.2.3 Summary

In conclusion, our experimental evaluations of the complex anomaly detection model have demonstrated its remarkable proficiency in detecting anomalous events within video sequences. The model has proven successful even in scenarios with subtle abnormalities that are difficult to identify visually. Despite the challenges associated with identifying anomalies in real-world applications, such as their infrequent occurrence, complexity, and unpredictability, our proposed model has shown significant capabilities in detecting multiple anomalous events within a single video.

The qualitative analysis, provides a visual representation of the model's performance. The frame anomaly scores generated by the model accurately identify anomalies, marked by red circles, while normal events are denoted by green circles. The model's ability to detect anomalies within complex video sequences, even in the presence of challenges and occasional incorrect detections (highlighted by yellow circles), highlights its potential in recognizing dangerous situations and abnormal events across various real-world applications.

The detailed analysis of each class from UCF-Crime dataset showcases the model's performance in detecting and differentiating between normal and anomalous activities. From scenarios involving abuse, arrest, arson, assault, burglary, explosion, fighting, road accidents, robbery, shootings, shoplifting, stealing, vandalism, to normal situations, the model consistently identifies anomalies, providing valuable insights in their occurrences.

Overall, our proposed anomaly detection model presents a promising solution for real-world applications that require the recognition of anomalous activities within video sequences. Its effectiveness in identifying anomalies, even in complex scenarios, makes it a valuable tool for enhancing security, safety, and surveillance systems. Further improvements and refinements can be pursued to enhance the model's accuracy and reliability, ensuring its robust performance across diverse contexts.

6.11.3 Analysis on The Web Dataset, GBA and ShanghaiTech

After analyzing examples from the different classes in the UCF-Crime Dataset, this section presents the results obtained in The Web Dataset, GBA dataset and ShangaiThech, without fine-tuning, to shows the generalization capability of the proposal in images different to those used for training.

6.11.3.1 The Web Dataset

First, Figure 6.24(a) from The Web Dataset shows a scene from Pamplona's 'Running Of The Bulls' event. In the beginning of the clip, people are running away from the bulls, which is a normal scenario for this event. Our model correctly recognizes it as a normal

event. However, when two bulls fall down and the person running looks back, our model detects this as abnormal behavior and marks it with a red circle. After that, when both bulls stand up, our model misclassifies it as normal behavior and marks it with a yellow circle. A few frames later, when the bulls start running in different directions, our model correctly detects it as an anomaly. On the other hand, Figure 6.24(b) shows a crowd fighting scenario. Throughout the clip, people are engaged in fighting with wood sticks. Our model predicts this behavior as abnormal throughout the entire duration of the clip.

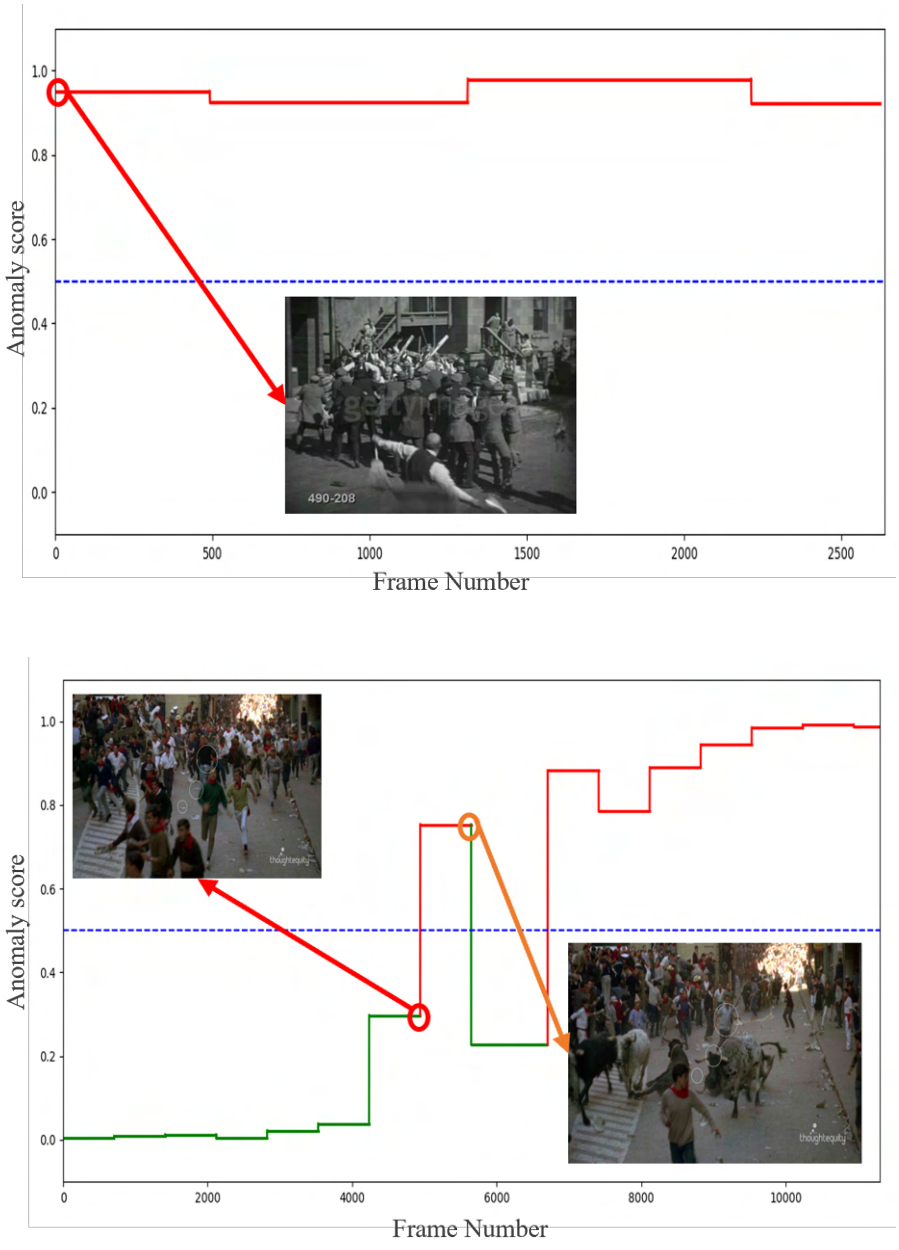


Figure 6.24: Qualitative visual result in The Web Dataset.

6.11.3.2 GBA Dataset

Figure 6.25 shows two examples from GBA dataset. It is a challenging dataset, that includes anomalous events in the wild. The example shown in Figure 6.25(a), depicts a person walking in a hallway. Suddenly, the person falls down on the floor, which our model predicts as an anomaly and marks it with a red circle. When the person stands up and continues walking, our model predicts it as normal behavior. However, when the person looks back at the camera, our model again predicts it as an anomaly.

Another example from GBA dataset is presented in Figure 6.25(b), that shows a stampede scenario. In the beginning, the hallway is empty, so our model predicts it as normal. Suddenly, a group of students enters the hallway in a frenzied manner, which is marked with a red circle, indicating an anomaly. When the students move out of the frame, our model predicts it as normal behavior. The yellow marked area represents a misclassification by the model.

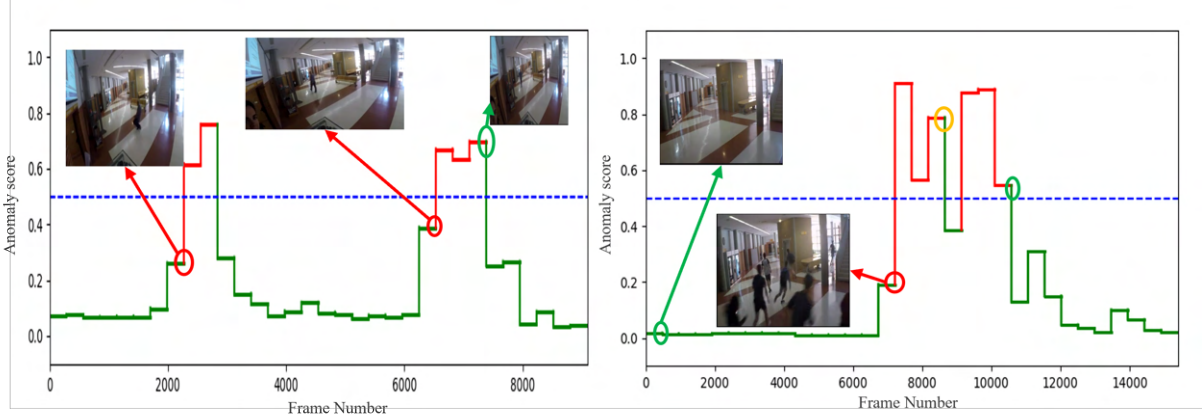


Figure 6.25: Qualitative visual result in GBA dataset.

6.11.3.3 ShanghaiTech Dataset

In Figure 6.26(a) from ShanghaiTech dataset, we can observe a scene where a group of students is walking on a campus. This is considered a normal activity, and our model correctly recognizes it as such, marking it with a green circle. However, there is a sudden change in the behavior of one of the students. Instead of continuing to walk like the others, this student starts jumping. This unexpected action deviates from the typical walking behavior and is considered an anomaly. Our model accurately detects this abnormal behavior and marks it with a red circle.

While, in Figure 6.26(b) we observe a group of students walking on a campus, which our model correctly identifies as a normal activity, marked with a green circle. However, at a certain point, one of the students exhibits an unexpected behavior. This student

throws their bag upward and jumps to catch it. This action deviates from the typical walking behavior, and our model identifies it as an anomaly, marking it with a red circle.

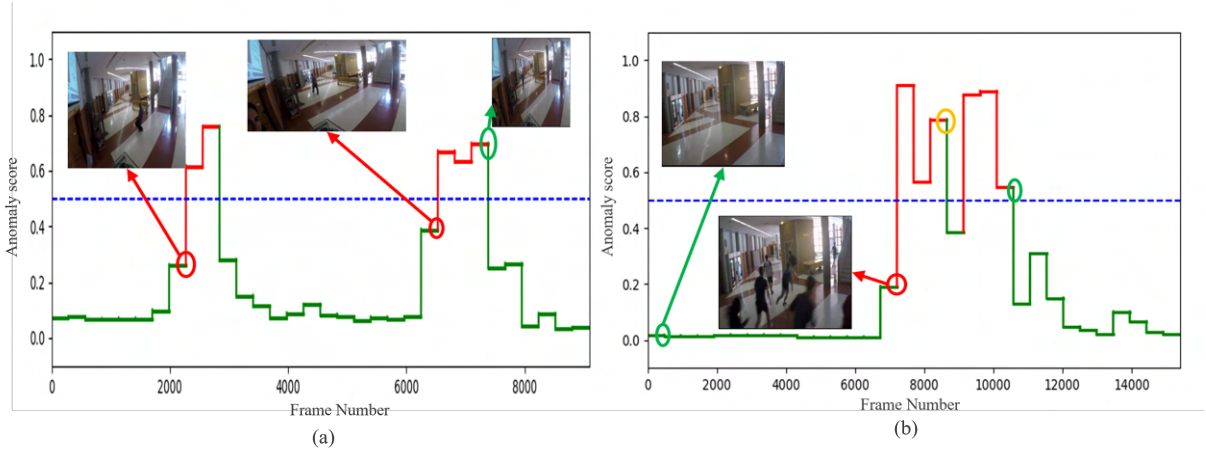


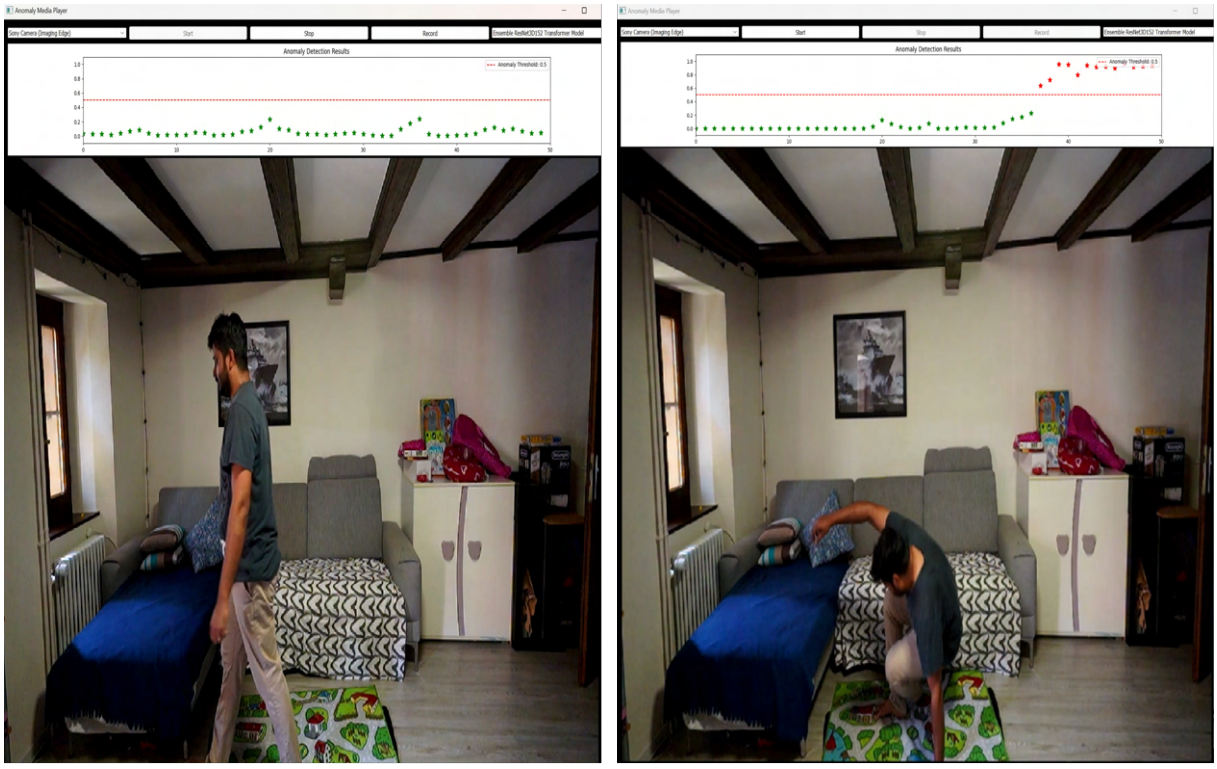
Figure 6.26: Qualitative visual result in ShanghaiTech dataset.

6.11.4 Analysis of the Real-time Anomaly Detection Prototype

As it has been explained in Section 6.7, we have developed a GUI for testing the final proposal in real-time applications.

The clips shown in Figure 6.27 depict scenarios captured from a real-time environment. Figure 6.27(a) illustrates a normal scenario where a person is walking inside a room. Our anomaly detection model successfully recognizes this event as normal.

In Figure 6.27(b), after a few frames, the person falls down, resulting in an abnormal event. This is identified as an anomaly by our model, which is evident from the stars (representing frames) appearing above the threshold limit and the color change to red. This detection occurs because the model has been trained to recognize deviations from typical human movements. A fall represents a significant divergence from standard movement patterns, thereby triggering the model to signal an anomaly due to its learned parameters differentiating between regular movements and those indicating potential emergencies.



(a) (b)

Figure 6.27: Real-time anomaly detection examples: (a) normal scene, (b) abnormal scene.

Chapter 7

Conclusions and Future Works

The weakly-supervised methodology for anomaly detection that has been presented in this research, which shows significant improvements over current state of the art techniques, is discussed in conclusion. Our method has demonstrated higher performance in detecting anomalies across many domains by providing an improved ranking loss function coupled with an attention mechanism for feature extraction. Quantitative and qualitative findings that show the proposed architecture's capacity to deliver precise and quick anomaly detection in real-world scenarios further support its real-time applicability.

7.1 Conclusions

The research presented in this PhD significantly advances the field of anomaly detection, particularly by demonstrating the effectiveness of a weakly-supervised approach. This strategy, rooted in the objectives outlined in section 1.2, demonstrates the potential of

obtaining reliable anomaly detection outcomes without the need for extensive and expensive annotations, presenting a more practical solution for organizations implementing anomaly detection technologies.

A central aspect of our study was the application of advanced anomaly detection techniques, as detailed in subsection 2.2.3. These methodologies, which include the use of A3DR, T3DR, and an E3DRT Models, were pivotal in efficiently monitoring activities in surveillance data and detecting object-related anomalies, thus fulfilling the objectives outlined in section 1.2 and adhering to the classifications proposed by Chandola et al. [17].

The significance of frame-level labeling in weakly supervised settings, which we emphasized in our research, is a direct outcome of the sophisticated anomaly detection algorithms we developed, as discussed in subsection 2.4.1. This approach has greatly enhanced the performance of our systems, providing a finer understanding of spatio-temporal features in surveillance scenarios.

Looking forward, we aim to expand upon this foundational work. Integrating multiple data sources, as indicated in section 1.2, holds promise for extending the reach and effectiveness of our anomaly detection methods. We have already observed this strategy's potential in preliminary studies with datasets like the UCF-Crime [5], ShanghaiTech [4], GBA [6] and The Web Dataset [54].

Furthermore, our investigation into advanced feature extraction processes along with ensemble modeling as described in our contributions (section 1.4), has opened new avenues for research. The E3DRT model, in particular, demonstrated superior performance, highlighting the value of combining multiple modeling approaches. Future research could delve into optimizing these models to further enhance the anomaly detection capabilities of weakly-supervised systems.

We hope this PhD thesis will inspire future research in weakly-supervised approaches for anomaly detection. As the demand for efficient and cost-effective systems grows, developing innovative methods to navigate the challenges of sparse or expensive annotations becomes increasingly important. Our work illustrates that through innovative loss functions and a blend of attention mechanisms and ensemble models, it is possible to achieve superior performance over current state-of-the-art techniques.

In summary, this research presents a novel, comprehensive approach to anomaly detection, demonstrating significant promise in both performance and real-world application. Our extensive evaluations across well-known datasets, coupled with the successful application in various scenarios, highlight the importance of sophisticated frame-level labeling and the integration of diverse data sources and methodologies in the progression of anomaly detection technologies.

7.2 Future Works

Building on the comprehensive research presented in this PhD thesis, there are several promising directions for future work. These areas are proposed with the intent of addressing current challenges and leveraging the foundations laid in the earlier chapters, particularly the methodologies in Chapter 6 and the results discussed in Chapter 6.8.

1. **Threshold Determination**: Investigate alternative methods for determining the threshold used to classify video segments as normal or anomalous. Adaptive thresholding or confidence-based thresholding techniques can be explored to make the threshold more robust to different levels of noise or ambiguity in the data.

2. **Temporal and Contextual Modeling** Incorporating advanced temporal modeling techniques, as discussed in subsection [subsec:anomaly-techniques], can better capture the temporal context and subtle nuances in video sequences. This approach can enhance the detection of complex anomalies, thus building upon the initial methodologies developed.

3. **Knowledge Representation**: Leverage prior knowledge or semantic information about scenes or objects in the videos to improve generalization and interpretability. Integrate knowledge representation or reasoning mechanisms, such as ontologies, graphs, or attention mechanisms, to enhance the system's performance.

4. **Collaborative and Multdisciplinary Approaches** The integration of insights from various disciplines, as indicated in the objectives (section 1.2), can enrich anomaly detection research. Collaborations with experts in psychology, criminology, or urban studies, for example, could offer new perspectives and enhance the applicability of the systems developed.

5. **New Dataset Creation**: Create a new CCTV surveillance anomaly detection dataset that addresses the shortcomings observed in existing datasets. Ensure the new dataset features more consistent video lengths, avoids dominance by a few videos, and contains fewer digitally edited images and scene cuts. This will provide a more realistic evaluation of anomaly detection methods in real-world scenarios.

6. **Optimization of Error Rates**: Investigate approaches to skew the error rates in a real-world scenario to obtain more false negatives and fewer false positives. Conduct additional tests to evaluate the system's ability to detect actual abnormalities, considering capturing at least one frame per abnormality.

7. **Adapting to Evolving Technologies and Trends** Finally, it is essential to keep pace with rapidly evolving technologies. This includes adapting anomaly detection methodologies to emerging video surveillance technologies and data formats, ensuring that the systems remain relevant and effective.

8. **Integration of Multiple Data Sources**: Investigate the combination of different data sources, such as audio, video, and sensor data, to improve anomaly detection capabilities enable comprehensive surveillance.

By pursuing these future research directions, we can further enhance the proposed weakly-supervised approach for anomaly detection in video surveillance. Advancements in threshold determination, temporal modeling, knowledge representation, dataset creation, exploration of different abnormalities, optimization of error rates, anomaly relevance research, and integration of multiple data sources will contribute to the development of more effective, affordable, and reliable anomaly detection systems.

Bibliography

[1] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, "T-c3d: Temporal convolutional 3d network for real-time action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[2] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[3] S. Islam, A. Dash, A. Seum, A. H. Raj, T. Hossain, and F. M. Shah, "Exploring video captioning techniques: A comprehensive survey on deep learning methods," *SN Computer Science*, vol. 2, no. 2, pp. 1–28, 2021.

[4] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.

[5] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[6] Geintra Group, "Gba dataset," https://www.geintra-uah.org/datasets/gba, 2021, last access: 15 Nov 2023.

[7] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan, "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2276–2280.

[8] E. Duman and O. A. Erdem, "Anomaly detection in videos using optical flow and convolutional autoencoder," *IEEE Access*, vol. 7, pp. 183 914–183 923, 2019.

[9] Y. Zhu and S. Newsam, "Motion-aware feature for improved video anomaly detection," *arXiv preprint arXiv:1907.10211*, 2019.

[10] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using oriented violent flows," *Image and vision computing*, vol. 48, pp. 37–41, 2016.

[11] S. Mohammadi, A. Perina, H. Kiani, and V. Murino, "Angry crowds: Detecting violent events in videos," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–18.

[12] M. Kass and A. Witkin, "Analyzing oriented patterns," *Computer vision, graphics, and image processing*, vol. 37, no. 3, pp. 362–385, 1987.

[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] J. Wang and A. Cherian, "Gods: Generalized one-class discriminative subspaces for anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8201–8211.

[16] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[17] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882

[18] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1257–1272, 2012.

[19] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for video Technology*, vol. 18, no. 11, pp. 1544–1554, 2008.

[20] A. C. Bahnsen, "Building ai applications using deep learning," *Blog Total Fraud Protection*, 2016.

[21] W. Zhang, J. Tanida, K. Itoh, and Y. Ichioka, "Shift-invariant pattern recognition neural network and its optical architecture," in *Proceedings of annual conference of the Japan Society of Applied Physics*. Montreal, CA, 1988, pp. 2147–2151.

[22] L. Medsker and L. C. Jain, *Recurrent neural networks: design and applications*. CRC press, 1999.

[23] M. I. Sarker, C. Losada-Gutiérrez, M. Marrón-Romera, D. Fuentes-Jiménez, and S. Luengo-Sánchez, "Semi-supervised anomaly detection in video-surveillance scenes in the wild," *Sensors*, vol. 21, no. 12, p. 3993, 2021.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[25] M. I. Sarker, M. Marrón-Romera, and C. Losada-Gutiérrez, "Real-time weakly supervised anomaly detection with attention mechanism," in *European Symposium on Computer and Communications (ESCC 2023)*, Manchester, UK, April 2023.

[26] A. Sanchez-Caballero, S. de López-Diz, D. Fuentes-Jimenez, C. Losada-Gutiérrez, M. Marrón-Romera, D. Casillas-Perez, and M. I. Sarker, "3DFCNN: Real-time action recognition using 3d deep neural networks with raw depth information," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 24 119–24 143, 2022.

[27] A. C. Cob-Parro, C. Losada-Gutiérrez, M. M. Romera, A. G. Vicente, I. B. Muñoz, and M. I. Sarker, "A proposal on stampede detection in real environments." in *IPINWiP*, 2021.

[28] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225–1234.

[29] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[30] E. H. Krause, "High altitude research with v-2 rockets," *Proceedings of the American Philosophical Society*, vol. 91, no. 5, pp. 430–446, 1947.

[31] "Pelco," 2023, accessed: 2023-06-06. [Online]. Available: https://en.wikipedia.org/wiki/Pelco

[32] "Advantages of surveillance cameras in schools," http://www.ehow.com/facts_5615866_advantages-surveillance-cameras-schools.html, accessed: 2023-06-06.

[33] R. Xue, J. Chen, and Y. Fang, "Real-time anomaly detection and feature analysis based on time series for surveillance video," in *2020 5th International Conference on Universal Village (UV)*. IEEE, 2020, pp. 1–7.

[34] J. Smith and K. Johnson, "Nyclu report: Surveillance camera increase in new york," *New York Community Board District Annual Report*, pp. 15–22, 2006.

[35] J. Doe, M. Smith, and K. Johnson, "Surveillance camera proliferation in spain: A comprehensive study," *Journal of Security and Surveillance Studies*, vol. 29, no. 2, pp. 65–79, 2023.

[36] T. C. Post, "Copenhagen police setting up more surveillance cameras," The Copenhagen Post, Tech. Rep., 2021. [Online]. Available: https://cphpost.dk/2021-05-11/news/copenhagen-police-setting-up-more-surveillance-cameras/

[37] S. M. of Interior, "Crime prevention effects of closed circuit television: a systematic review," Spanish Ministry of Interior, Tech. Rep., 2023.

[38] E. País, "Marbella's smart surveillance system is watching you," 2019. [Online]. Available: https://english.elpais.com/elpais/2019/11/27/inenglish/1574849134_892168.html

[39] G. Pereira and C. Raetzsch, "From banal surveillance to function creep: Automated license plate recognition (alpr) in denmark," *Surveillance & Society*, vol. 20, no. 3, 2022.

[40] Pinkerton, "Cctv in criminal investigations," *Pinkerton Insights*, 2022. [Online]. Available: https://www.pinkerton.com/our-insights/blog/cctv-in-criminal-investigations

[41] R. Baran, T. Rusc, and P. Fornalski, "A smart camera for the surveillance of vehicles in intelligent transportation systems," *Multimedia Tools and Applications*, vol. 75, pp. 10 471–10 493, 2016.

[42] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson, 2016.

[43] C. C. Loy, T. Xiang, and S. Gong, "Activity understanding in video surveillance," in *Computer Vision and Image Understanding*, vol. 115, no. 3. Elsevier, 2010, pp. 420–430.

[44] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[45] D. M. Hawkins, *Identification of Outliers*. Springer Dordrecht, 1980.

[46] DataDx, "Applying anomaly detection to healthcare data," 2020. [Online]. Available: https://www.datadx.com/applying-anomaly-detection-to-healthcare-data/

[47] T. Pourhabibi, K.-L. Ong, B. H. Kam, and Y. L. Boo, "Fraud detection: A systematic literature review of graph-based anomaly detection approaches," *Decision Support Systems*, vol. 133, p. 113303, 2020.

[48] Z. Yang, X. Liu, T. Li, D. Wu, J. Wang, Y. Zhao, and H. Han, "A systematic literature review of methods and datasets for anomaly-based network intrusion detection," *Computers & Security*, vol. 116, p. 102675, 2022.

[49] D. R. Patrikar and M. R. Parate, "Anomaly detection using edge computing in video surveillance system," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 2, pp. 85–110, 2022.

[50] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016.

[51] M. J. V. Leach, "Automatic human behaviour anomaly detection in surveillance video," Ph.D. dissertation, Heriot-Watt University, 2015.

[52] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep learning strong parts for pedestrian detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1904–1912.

[53] ManageEngine, "Detecting anomalies - the what, the why & the how," https://www.manageengine.com/log-management/ueba/resources/detecting-anomalies-the-what-the-why-the-how.html, accessed: May 2, 2024.

[54] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.

[55] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Holistic features for real-time crowd behaviour anomaly detection," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 918–922.

[56] J. Wang and Z. Xu, "Spatio-temporal texture modelling for real-time crowd anomaly detection," *Computer Vision and Image Understanding*, vol. 144, pp. 177–187, 2016.

[57] A. A. Sodemann, M. P. Ross, and B. J. Borghetti, "A review of anomaly detection in automated surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1257–1272, 2012.

[58] A. A. Ghorbani, W. Lu, M. Tavallaee, A. A. Ghorbani, W. Lu, and M. Tavallaee, "Theoretical foundation of detection," *Network Intrusion Detection and Prevention: Concepts and Techniques*, pp. 73–114, 2010.

[59] B. Benfold, "The acquisition of coarse gaze estimates in visual surveillance," Ph.D. dissertation, University of Oxford, 2011. [Online]. Available: http://www.robots.ox.ac.uk/ActiveVision/Publications/benfold_dphil2011/benfold_dphil2011.html

[60] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[61] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1-2, pp. 31–71, 1997.

[62] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.

[63] F. Muhlenbach, S. Lallich, and D. A. Zighed, "Identifying and handling mislabelled instances," *Journal of Intelligent Information Systems*, vol. 22, no. 1, pp. 89–109, 2004.

[64] D. C. Brabham, "Crowdsourcing as a model for problem solving: An introduction and cases," *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.

[65] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds." *Journal of machine learning research*, vol. 11, no. 4, 2010.

[66] R. Urner, S. B. David, and O. Shamir, "Learning from weak teachers," in *Artificial intelligence and statistics*. PMLR, 2012, pp. 1252–1260.

[67] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[68] S. Das, B. Amoedo, F. De la Torre, and J. Hodgins, "Detecting parkinsons' symptoms in uncontrolled home environments: A multiple instance learning approach," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 3688–3691.

[69] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Advances in neural information processing systems*, vol. 15, 2002.

[70] J. Davis, V. Santos Costa, S. Ray, and D. Page, "Tightly integrating relational learning and multiple-instance regression for real-valued drug activity prediction," in *Proceedings of the 24th International Conference on Machine Learning, ICML*, vol. 7, 2007, pp. 425–432.

[71] Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic, and S. Vucetic, "Aerosol optical depth prediction from satellite observations by multiple instance regression," in *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 2008, pp. 165–176.

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

[73] C. Bergeron, G. Moore, J. Zaretzki, C. M. Breneman, and K. P. Bennett, "Fast bundle algorithm for multiple-instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1068–1079, 2011.

[74] O. Maron and A. W. Moore, "The multiple instance problem in learning for natural language processing," in *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*, 1998, pp. 271–277.

[75] X. Zhang, X. Xu, J. Yang, Z. Zhang, S. Yan *et al.*, "Multiple instance learning for face detection," in *European Conference on Computer Vision*, 2014, pp. 179–192.

[76] P. Wu, J. Liu, and F. Shen, "A deep one-class neural network for anomalous event detection in complex scenes," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 7, pp. 2609–2622, 2019.

[77] J. Qin and S. Belongie, "Multiple instance learning for cluttered scene classification," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016, pp. 1256–1265.

[78] W. Wang, N. Carlini, and M. Dudik, "Learning to prevent leakage in machine learning models," in *International Conference on Machine Learning*, 2019, pp. 6554–6563.

[79] W. Li, W. Zhang, D. Xu, and X. Zhang, "Multiple instance ranking," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016, pp. 2480–2489.

[80] D. Zhang, F. Wang, L. Si, and T. Li, "Maximum margin multiple instance clustering with applications to image and text clustering," *Ieee transactions on neural networks*, vol. 22, no. 5, pp. 739–751, 2011.

[81] M.-L. Zhang and Z.-H. Zhou, "Multi-instance clustering with applications to multi-instance prediction," *Applied intelligence*, vol. 31, no. 1, pp. 47–68, 2009.

[82] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, vol. 10, 1997.

[83] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & security*, vol. 28, no. 1-2, pp. 18–28, 2009.

[84] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, 2010.

[85] M. Nawaz and J. Ahmed, "Cloud-based healthcare framework for real-time anomaly detection and classification of 1-d ecg signals," *PLoS One*, vol. 17, no. 12, p. e0279305, 2022.

[86] R. Isermann, "Fault-diagnosis systems: An introduction from fault detection to fault tolerance," *Springer Science & Business Media*, 2006.

[87] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pp. 170–180, 2002.

[88] J.-Y. Chen, Z. Gan, and J. Xiao, "Outlier detection with autoencoder ensembles," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 1638–1646.

[89] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 665–674, 2017.

[90] W. Lu, Y. Cheng, C. Xiao, S. Chang, S. Huang, B. Liang, and T. Huang, "Unsupervised sequential outlier detection with deep architectures," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4321–4330, 2017.

[91] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," *Proceedings of the International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 146–157, 2017.

[92] J. Zhao, M. Mathieu, and Y. LeCun, "Energy-based generative adversarial network," *arXiv preprint arXiv:1609.03126*, 2016.

[93] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 481–490.

[94] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," *Advances in neural information processing systems*, vol. 31, 2018.

[95] S. Wang, Y. Zeng, X. Liu, E. Zhu, J. Yin, C. Xu, and M. Kloft, "Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network," *Advances in neural information processing systems*, vol. 32, 2019.

[96] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.

[97] H. Wang, G. Pang, C. Shen, and C. Ma, "Unsupervised representation learning by predicting random distances," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI'20, 2021.

[98] M.-N. Nguyen and N. A. Vien, "Scalable and interpretable one-class svms with deep learning and random fourier features," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 2019, pp. 157–172.

[99] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *ICML*, 2016, pp. 478–487.

[100] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," *ICLR*, 2018.

[101] G. Pang, C. Yan, C. Shen, A. van den Hengel, and X. Bai, "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *CVPR*, 2020, pp. 12 173–12 182.

[102] S. Fan, C. Shi, and X. Wang, "Abnormal event detection via heterogeneous information network embedding," in *CIKM*, 2018, pp. 1483–1486.

[103] G. Pang, C. Shen, and A. van den Hengel, "Deep anomaly detection with deviation networks," in *KDD*, 2019, pp. 353–362.

[104] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *CVPR*, 2018, pp. 3379–3388.

[105] P. Zheng, S. Yuan, X. Wu, J. Li, and A. Lu, "One-class adversarial nets for fraud detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1286–1293.

[106] C. P. Ngo, A. A. Winarto, C. K. K. Li, S. Park, F. Akram, and H. K. Lee, "Fence gan: towards better anomaly detection," *arXiv preprint arXiv:1904.01209*, 2019.

[107] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2898–2906.

[108] M. Bertini, A. Del Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 320–329, 2012.

[109] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.

[110] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 341–349.

[111] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X.-S. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1933–1941.

[112] R. Nawaratne, D. Alahakoon, D. De Silva, and X. Yu, "Spatiotemporal anomaly detection using deep learning for real-time video surveillance," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 393–402, 2019.

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

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

[115] C. He, J. Shao, and J. Sun, "An anomaly-introduced learning method for abnormal event detection," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29 573–29 588, 2018.

[116] J. Zhang, L. Qing, and J. Miao, "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4030–4034.

[117] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1237–1246.

[118] S. Tariq, H. Farooq, A. Jaleel, S. M. Wasif *et al.*, "Anomaly detection with particle filtering for online video surveillance," *IEEE Access*, vol. 9, pp. 19 457–19 468, 2021.

[119] G. Yu, S. Wang, Z. Cai, E. Zhu, C. Xu, J. Yin, and M. Kloft, "Cloze test helps: Effective video anomaly detection via learning to complete video events," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 583–591.

[120] P. López Miguel, "Detección de actividades anómalas en espacios públicos mediante redes neuronales profundas," Master's thesis, Universidad de Alcalá. Escuela Politécnica Superior, 2019.

[121] M. Baptista-Ríos, C. Martínez-García, C. Losada-Gutiérrez, and M. Marrón-Romera, "Human activity monitoring for falling detection. a realistic framework," in *2016 International conference on indoor positioning and indoor navigation (IPIN)*. IEEE, 2016, pp. 1–7.

[122] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[123] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[124] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[125] C. Dyer, "Notes on adagrad," *School of Computer Science, Carnegie Mellon University*, vol. 5000, p. 15, 2013.

[126] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[127] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[128] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[129] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

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

[131] Z. Li, T. Lin, X. Shang, and C. Wu, "Revisiting weighted aggregation in federated learning with neural networks," *arXiv preprint arXiv:2302.10911*, 2023.

[132] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[133] S. Hochreiter, "The vanishing gradient problem for recurrent nets and solutions," *Technical Report. Institut f. Informatik, Technische Univ. Munich*, 1991.

[134] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition*, 2017, pp. 3154–3160.

[135] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," *arXiv preprint arXiv:1711.10305*, 2017.

[136] M. A. Ganaie, M. Hu, A. Malik, M. Tanveer, and P. Suganthan, "Ensemble deep learning: A review," *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 2022.