Adversarial Robustness in Video Surveillance: A GAN-Based Attack Generation and Defence Framework for YOLO

Saeed Matar Aljaberi

A thesis submitted in partial fulfilment of the requirements of University of Staffordshire for the degree of Doctor of Philosophy

University of Staffordshire
School of Digital, Technology, Innovation and Business
United Kingdom

October 2025

Abstract

The operational integrity of Artificial Intelligence (AI)-powered video surveillance systems is critically threatened by adversarial attacks that exploit vulnerabilities in object detectors like YOLO (You Only Look Once). This research proposes a comprehensive dual-framework to both assess and mitigate this threat. On the offensive front, we develop an enhanced Generative Adversarial Neural Network (GAN) attack model, incorporating a novel composite loss function that combines adversarial, L1 perceptual, and cosine similarity losses. This architecture forces the generator to produce adversarial examples that are not only potent in evading detection but also semantically coherent and realistic. Defensively, we fortify the YOLO object detector by integrating a Tracking-Learning-Detection (TLD) module, creating a YOLO-TLD framework that enhances resilience through robust long-term tracking and online P-N learning, which continuously updates the detector based on tracking consistency and error correction.

The proposed offensive and defensive models were rigorously evaluated against each other using benchmark datasets, including COCO, VOC 2007, and the realistic VIRAT surveillance video dataset. The results demonstrate a critical security arms race: the enhanced GANN model achieved a remarkable fooling rate of over 92% on static images and 81% on video sequences, effectively compromising a standard YOLO detector. Conversely, the defensive YOLO-TLD system showed significant resilience, raising detection accuracy on the COCO dataset under adversarial conditions from 85% to 90.5%. However, this defense was not absolute; when subjected to the most sophisticated GANN attacks, the performance of even the fortified YOLOTLD experienced a dramatic decline, with accuracy in certain scenarios plummeting from 92% to less than 5%.

These findings highlight the severe and practical threat of GAN-based attacks while validating the value of integrated tracking and learning for defense. The study concludes that a fundamental shift towards adversarial training and hybrid, tamper resistant architectures is imperative. The contributions of this work provide a critical

methodological framework and benchmark for developing next-generation surveillance systems capable of withstanding evolving adversarial threats.

Acknowledgements

First and foremost, I am deeply grateful to Allah for granting me the strength, resilience, and inspiration to overcome the many challenges I encountered throughout this academic journey. To my beloved parents, whose unwavering support and guidance have been my foundation, I extend my heartfelt gratitude. Their love, encouragement, and assistance (especially in caring for my children, Maryam and Mater) allowed me to pursue my educational aspirations with great focus and peace of mind. Their selfless devotion and understanding have been a constant source of strength that made the completion of this work possible.

My deepest appreciation also goes to my dear wife, whose patience, steadfast support, and encouragement have accompanied me through every stage of this journey. Her remarkable understanding and sacrifices made it possible for me to persevere and succeed. I owe her a debt of gratitude that words can hardly express. Her faith in me and in the purpose of this pursuit has been unshakeable.

I also wish to express my profound thanks to my esteemed supervisors, Dr. Asma Patel and Dr. Ahmed N. Al-Masri, for their invaluable expertise, insightful guidance, and dedicated mentorship. Their thoughtful feedback, constructive criticism, and continuous encouragement have been instrumental in shaping the direction and outcomes of this study. Their belief in my abilities has motivated me to strive for excellence and to overcome every obstacle with confidence.

I am equally thankful to all faculty members who, in various ways, enriched my academic experience. Each professor’s commitment to their students, their teaching of how knowledge is best conveyed, and their ability to challenge and expand my thinking have left an enduring mark on my intellectual and personal development.

Finally, I extend my sincere appreciation to my extended family, friends, and well-wishers for their unwavering faith, encouragement, and kindness. Their trust

and positive words have lifted me in moments of difficulty, reminding me of the strength found in community and the shared journey of perseverance toward achieving one’s goals.

Declaration of Authorship

I, Saeed Matar Al Jaberi, declare that this thesis titled “Adversarial Robustness in Video Surveillance: A GAN-Based Attack Generation and Defence Framework for YOLO” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly whilst in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed.
- This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Signed:

Date: 12/12/2025

Publications

Al Jaberi et al. (2023). Object tracking and detection techniques under GAN threats: a systemic review. Applied Soft Computing, 139, p.110224.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial intelligence
ANN	Adversarial neural network
BFGS	Broyden–Flecther–Goldfarb–Shanno method
CNN	Convolutional neural network
CW	Carlini and Wagner method
CCTV	Closed circuit television
CycleGAN	Cycle-consistent adversarial networks
COCO	Common objects in context
CRNs	Cognitive radio networks
DL	Deep learning
DNN	Distributed neural networks
DDos	Distributed denial of service
DCGAN	Deep convolutional generative adversarial network
ELM	Extreme learning machine
FGSM	Fast gradient sign method
FGPA	Field-programmable gate array
GAN	Generative adversarial network
GAN	Generative adversarial neural network
GCC	Gulf Cooperation Council
HOG	Histogram of oriented gradients
HOD	Head of department
IoT	Internet of Things
KCF	Kernelised correlation filter
KLT	Kanade–Lucas–Tomasi
LBP	Local binary pattern
LSTM	Long short-term memory
L-BFGS	L-Broyden–Fletcher–Goldfarb–Shanno
ML	Machine learning
MSE	Mean squared error
MAE	Mean absolute error
MOT	Multi-object tracking
MTGAN	Multi-task generative adversarial network
MMLS	Mask mean loss function
NN	Neural network
NM-GANs	Noise-modulated generative adversarial network
RNN	Recurrent neural network
R-CNN	Region-based convolutional neural network
RPN	Regional proposed network
RAP	Regional average pooling

ROI	Region of interest
PEP	Probability elastic part
PUs	Primary users
PASCAL	Visual object classes
SUs	Secondary users
SIFT	Scale-invariant feature transform
SLR	Systematic literature review
SVM	Support vector machine
SRC	Sparse representation-based classification
SORT	Simple online and real-time tracking
TLD	Tracking–learning Detection
TPI	Toxicity probability interval
TRA	Telecommunication Regulatory Authority
UAVs	Unmanned aerial vehicles
UAE	United Arab Emirates
VOCs	Visual object classes
YOLO	You Only Look Once

Chapter 1

Introduction

This chapter addresses the challenges posed by adversarial attacks to surveillance systems, for which robust security measures and research are necessary for defence. Moreover, it presents the problem statement briefly, highlighting the potential power of attacks from phishing exploits, deepfake generators, and generative adversarial networks (GANs) that can deceive artificial intelligence (AI)-powered surveillance systems. The research addresses vulnerabilities in surveillance systems and proposes innovative approaches to enhance defence mechanisms and counter offensive strategies effectively. The research's aims and objectives include assessing advanced techniques, strengthening attack methods and evaluating the criteria mechanisms. The research questions are posed to guide the study of the performance of AI-based classifiers and the effect of GAN-generated mock data. The research framework is built in consideration of data samples and computer constraints, and ethical issues are always at the forefront of the approach. Furthermore, the context is provided to readers, allowing them to draw their own conclusions and develop recommendations to address this problem, which still exists in surveillance systems.

1.1. Overview

Adversarial attacks are used increasingly in manipulating surveillance applications and video data, highlighting the importance of developing robust security measures and the need for research on improving techniques and approaches for defending against these attacks (Nguyen et al., 2023a). One such technique used in adversarial attacks on surveillance applications is the generation of negative samples. Attackers can make small changes to the input data, such as modifying an image or video, that are invisible to humans but can cause an AI-based model to misclassify or fail to detect specific objects or behaviours. For example, an attacker could generate an adversarial example of a person wearing a mask undetectable to a surveillance camera's mask recognition algorithm, allowing them to evade detection. Another technique (Khaleel et al., 2024) involves using GANs to generate synthetic data that can be used to deceive machine learning (ML) models.

Zhang et al. (2020a) stated that attackers can train GANs on a dataset of actual surveillance footage and use them to generate synthetic videos that appear to show normal behaviour but contain malicious activity. For example, an attacker could create a fake video of a person breaking into a building to confuse the surveillance system. Several techniques have been developed to defend against adversarial attacks on surveillance applications and video data. One approach is to train ML models on standard and adversarial examples to learn to recognise and reject negative input (Al-Garadi et al., 2020). This approach can improve the robustness of ML models and make them increasingly resilient to adversarial attacks.

Many approaches use learning processes that include context-aware rewarding functions, as highlighted in the study of Truong et al. (2020a). A system's robustness can be enhanced, and its resilience to malicious attacks can be increased by applying AI techniques, such as anomaly detection and training, on standard and adversarial examples. However, developing robust AI solutions for video surveillance in autonomous vehicles is an ongoing area of research. As attackers continue to find new ways to evade detection, researchers and practitioners must remain vigilant and continue to develop new and innovative approaches to enhance the security and reliability of these systems.

Video surveillance systems are everywhere. Common adversarial attacks on video surveillance include performing code injection, exfiltration of information and preventing access by flooding. Adversarial attacks on surveillance applications and video data are becoming increasingly sophisticated and can take many forms. For example, attackers can orchestrate distributed denial of service attacks to disrupt the network or use preliminary scanning and reconnaissance attacks followed by adversarial ML-based techniques, as highlighted by Kalbo et al. (2020).

These attacks can compromise the security and reliability of surveillance systems, potentially leading to safety issues and compromising privacy. Hence, in this research, we establish an evaluation framework for evaluating the extent of GAN-based attacks (generating fake images and videos) against ML-empowered video surveillance systems; the framework uses the You Only Live Once (YOLO) classifier (Zhang et al., 2022; Li & Cai, 2020; Yang et al., 2021; Su et al., 2022), which is considered one of the best of its kind. In this framework, we propose improving the YOLO system (in terms of its object-tracking capability) to increase its classification effectiveness. Meanwhile, we suggest improving the GAN-based system to fool the enhanced defensive YOLO classifier.

1.2. Problem Statement

The use of AI- and ML-based techniques to fortify video surveillance systems against attacks that target their learning systems and classifiers is increasing (Yasdinejad et al., 2022; Lan et al., 2018; Huang et al., 2018b; C'orovic´ et al., 2018). However, adversarial attacks are also diversifying their AI/ML techniques to evade the detection and identification mechanisms of these defensive systems (Liu et al., 2020; Ting et al., 2021; Wang et al., 2018; Lin et al., 2018).

GANs are amongst the most potent candidates for generating fake and illusory images and videos, which seriously threaten AI-augmented video surveillance systems. Malicious actors could use these techniques to launch actual, devastating attacks against defensive systems. Many researchers have addressed the ongoing virtual battle between GAN attacks and AI-based video surveillance defence systems (Suthishni & Kumar, 2022; Wu et al., 2021; Aung et al., 2021; Adarsh et al., 2020). These studies aimed to explore the deceiving strength of these proposed variants of GAN-generated attacks and assess their potential threats to AI-based video surveillance systems. The current study contributes to this important research area by exposing the ever-increasing threat of GAN-based fooling attacks, their versatility and their ability to deceive video surveillance systems and jeopardise security.

1.3. Research Questions

This research is guided by the following questions, which are designed to systematically address the vulnerabilities of AI-powered video surveillance systems to adversarial attacks. The objectives outline the specific steps taken to answer these questions. The research questions are as follows:

RQ1: Attack Generation: How can GANs be enhanced to generate potent and realistic adversarial examples that effectively deceive AI-based object detectors in video surveillance?

RQ2: Attack Assessment: What is the quantitative effect of these enhanced GAN-based attacks on the performance and reliability of a standard YOLO-based surveillance system?

RQ3: Defence Development: How can the resilience of the YOLO detector against sophisticated adversarial attacks be improved through integration with a tracking–learning–detection (TLD) module?

RQ4: Defence Evaluation: To what extent does the fortified YOLO-TLD defence system mitigate the influence of enhanced GAN-based attacks and maintain detection accuracy?

RQ5: Synthesis: What overarching principles and practical recommendations can be derived from the interaction between advanced attacks and defences to guide the development of robust future surveillance systems?

1.4. Aim and Objectives

This research is driven by a clear investigative pathway aimed at advancing the security of AI-powered video surveillance systems through a comprehensive analysis of adversarial attacks and defenses. The principal aim and corresponding objectives are structured as follows:

Principal Aim: To advance the security of AI-powered video surveillance systems through a comprehensive analysis of adversarial attacks and defenses. This aim is operationalized through five specific research objectives that structure the dissertation's investigative pathway from attack development to defense evaluation:

1. **Objective 1:** To develop an enhanced Generative Adversarial Network (GAN) framework capable of generating high-quality, realistic adversarial images and videos that effectively deceive AI-based object detectors. *(Addresses RQ1)*

2. **Objective 2:** To rigorously evaluate the impact of these enhanced GAN-based attacks on the performance and reliability of a standard YOLO-based video surveillance system, quantifying the severity of the threat. *(Addresses RQ2)*

3. **Objective 3:** To design and implement a fortified defensive system by enhancing the YOLO detector with a Tracking-Learning-Detection (TLD) module to improve its resilience against sophisticated adversarial attacks. *(Addresses RQ3)*
4. **Objective 4:** To assess the effectiveness of the fortified YOLO-TLD defense against the enhanced GAN-based attacks, measuring its ability to maintain detection accuracy and mitigate compromises. *(Addresses RQ4)*

5. **Objective 5:** To synthesize the findings from this adversarial interaction into a set of conclusions and practical recommendations for developing more robust, attack-resistant video surveillance systems in the future. *(Addresses RQ5)*.

1.5. Research Contributions

This research provides four novel contributions.

1. **Novel Adversarial Loss Function for GANs:** Development of a new loss function that enhances the deceptive capability of GANs. This innovation substantially increases the deception rate of generated adversarial examples, achieving a rate of up to 92% for images and over 81% on average for videos even when tested against a fortified detection system.
2. **Hybrid YOLO-TLD Defence Architecture:** Proposal and implementation of a novel defensive framework that integrates a TLD module with the YOLO classifier. This integration enhances temporal consistency and robustness,

resulting in a measurable improvement in detection accuracy (approximately 4%–5.5%) and establishing a resilient baseline against sophisticated attacks.

3. **Comprehensive Adversarial Evaluation Benchmark:** Establishment of a holistic evaluation methodology that rigorously tests the interaction between state-of-the-art attacks and defences under realistic conditions. This framework provides critical insights into the residual vulnerabilities of AI systems, quantifying the persistent threat landscape even when defences are improved.
4. **Synthesis for Secure AI Surveillance:** Aside from technical implementations, this work provides a critical analysis and synthesis of the adversarial cycle, leading to practical recommendations for building robust, next-generation video surveillance systems resistant to evolving threats.

Notably, building a promising theoretical and practical tamper-proof solution against adversarial attacks, especially those that are AI-empowered, on the basis of past literature is difficult (Robert & Vidya, 2022; Rawal & Manogaran, 2021; Dong et al., 2018; Taher et al., 2022). This study further confirms this conjecture to be ill, showing the power of GAN-based attacks against AI-empowered video surveillance systems and highlighting the race condition of defensive and offensive systems using diversified advanced AI-based optimisation techniques. With regard to the data samples, we use standard datasets that are commonly adopted to test images and videos of GAN-generated attacks. The expected outcomes of this research are as follows:

- The effectiveness of GAN-based attacks and the achievable success rates in fooling real-time detection systems will be demonstrated for the basic GAN system, and a proposed enhanced version of GAN will be used for the simulated attacks.
- On the defensive side, current YOLO-empowered systems for detecting fake images and videos generated by enhanced GAN attacks will be evaluated and tested.
- The lessons learned and appropriate useful recommendations will be provided.

1.6. Research Scope

Adversarial patterns show that ML algorithms can be fractured in unexpected ways that will jeopardise their effectiveness (Kumar et al., 2023; Ting et al., 2021; Wang et al., 2018; Lin et al., 2018). Thus, real-time video surveillance systems may not be able to detect all adversarial attacks they should catch, failing to qualify as tamper-proof systems. Any proposed video detection algorithm should address the problem of bias associated with neural networks, such as failing to detect dark-skinned faces accurately at night. In this context, lighting conditions are often overlooked, but they can substantially affect the effectiveness of cameras in detecting dark-skinned faces in low-light conditions.

The scope of our research is limited to instances of data samples fed during training by basic GAN or an improved version of it. The defensive system (YOLO based) specifically deals with the generated samples (images or videos) of the simulated GAN attack scenarios.

In terms of limitations, the results of this study's experiments are relative and sensitive to the computational power of the system. Thus, this study serves as proof of concept for GAN-based fooling attacks and for testing the scope of standard established datasets (images and video processing) used for GAN testing.

1.7. Ethical Considerations

This study involves several ethical considerations. Firstly, permission to conduct and continue this research must be obtained from the department heads and relevant officials of Staffordshire University. Secondly, the collected data should be autonomous and secure. In the case of practical implementation or real data collection, an agreement will be signed and sealed amongst the stakeholders. Finally, the researcher will ensure that the research findings are reliable and authentic, avoiding bias throughout the research process. Ethical approval for the research content has been obtained from the university, and the researcher asserts that the research findings are reliable and authentic.

1.8. Thesis Organisation

The layout of the thesis is structured as follows:

Chapter 1: Presents the thesis by stressing the rise of and the danger posed by adversarial attacks to video surveillance systems and the need for robust security and research to build defence layers against these threats. The chapter provides the problem statement, research objective, research questions and ethical considerations for establishing conclusions and recommendations.

Chapter 2: Presents a literature review of the state of the art in traditional ML algorithms and the research on emerging applications that use GAN in object tracking and detection for video surveillance. It also highlights the urgency of the need for research projects and technology to implement tamper-proof solutions and prevent adversarial threats in the future.

Chapter 3: Presents a strategic framework for GAN-based attacks on video monitoring systems and how to defend against them by understanding the capabilities of GAN and improving the identification methods to protect the systems.

Chapter 4: Introduces the GAN modelling technique for defence against adversarial attacks. This study emphasises the introduction of a novel architecture in object detection and tracking for integration with a defence system to enhance the attack models and how the generators and discriminators are on the saga machine.

Chapter 5: Focuses on the practical implementation and evaluation of the concepts presented in the preceding chapters. It presents the datasets and technical frameworks, the application of enhanced GAN models for generating synthetic data and the aftermath of primary video monitoring systems through hacking and defence. It aims to evaluate the detection capabilities after an obfuscated attack.

Chapter 6: Presents a comprehensive performance evaluation of the enhanced GAN-based attack model and YOLO-TLD defence system, focusing on detection accuracy, fooling rate and computational efficiency. Through empirical analysis of benchmark datasets, the chapter assesses the resilience of the AI-empowered surveillance framework against adversarial attacks. The results underscore the model's defensive robustness and GAN's deceptive strength, offering insights into the practical effectiveness and limitations of the proposed techniques.

Chapter 7: Summarises the conclusions and research perspectives, including the limitations of existing defences and practical ways to improve the natural resilience of systems through research and exploration of new datasets, adversarial training, model architectures and robustness.

Chapter 2

Literature Review

Chapter 2 explores the state of the art for traditional ML algorithms typically used in object tracking and detection and the pressing need to explore diverse types of GANs to address the increasing demand for AI applications. The chapter presents a literature review of traditional ML algorithms used in object tracking and detection. It emphasises the need to explore different types of GANs and their diverse use in object generation and detection and in the context of adversarial attacks. The primary focus points of this chapter are as follows:

- Review of literature on state-of-the-art object detection and tracking algorithms;
- Summary of studies on adversarial attacks and defences grouped into GAN- and non-GAN-related tasks;
- Discussion of the challenges, gaps and research directions in this field.

In line with the research objectives outlined in Chapter 1, this chapter addresses the following research question: What state-of-the-art AI-based object classifiers are implemented in video surveillance systems, and what are their weaknesses? The findings presented in this chapter will lay the foundation for exploring advanced techniques to enhance the resilience of surveillance systems against adversarial attacks in the subsequent chapters.

2.1 Introduction

Object tracking and detection technologies, along with their applications, have experienced exponential growth in the last decade. The increased research on object tracking and detection is driven by its diverse applications, such as video surveillance, human–machine interactions, traffic surveillance and malicious object or human behaviour detection. Malicious cyber, criminal and adversarial attacks on AI-based applications are increasing daily, threatening the security and safety of institutions, cities and companies worldwide. For instance, Abu Dhabi and

Dubai in the United Arab Emirates (UAE) faced around 86 cyberattacks at the beginning of 2018 (Chandra et al., 2019). These attacks severely affected AI-based surveillance systems because the data of over 14 million customers were leaked and exposed on the Internet. Malicious and adversarial attacks not only disrupt systems, business operations and services to citizens but also pose a threat to the economy and national security. Current advances in video surveillance require robust and tamper-proof object tracking and the application of detection techniques. The problem is that current applications and techniques are increasingly faced with common adversarial attacks, intrusions and hacks in unprecedented ways. Furthermore, super-fast image processing during real-time object tracking has become necessary for detecting malicious objects during surveillance.

Security prerequisites are becoming increasingly demanding in terms of reliability, availability and autonomous operation (not relying on human actors). Thus, efficient object tracking and detection techniques for security systems require independence from the human factor or need automated security measures, such as adopting ML algorithms to help in monitoring tasks. ML techniques enable feature learning and image/object generation and are critical for ensuring the high-order security and robustness of the deployed algorithms. Several challenges faced by video surveillance systems, including occlusion, viewpoint variations and illumination problems, are discussed in the latter part of this thesis to guide the research in this field.

The primary emphasis of this study is to examine the vulnerabilities of object tracking and detection techniques to GAN threats. It discusses various types of GAN threats and examines distinct adversarial attacks, such as black-box, white-box and grey-box attacks, on the basis of the threat model. This chapter systematically reviews object tracking and detection techniques under GAN threats and includes implications and concluding remarks for future challenges.

Nawaratne et al. (2020) investigated the possible advantages and obstacles of deploying intelligent video-based surveillance systems. They emphasised the growing need for intelligent surveillance systems to independently analyse and interpret video information to identify and mitigate potential security risks. The research explored the challenges faced when such systems are implemented, including the complexity of managing large amounts of real-time video data, the need for effective storage and retrieval methods and the need to address privacy and ethical concerns associated with surveillance technologies. The authors theoretically stressed the need for ongoing research and development to fully leverage image processing techniques and advance surveillance systems.

Kumari (2024) investigated the possible advantages and obstacles related to deploying intelligent video-based surveillance systems. They underscored the growing need for intelligent surveillance systems to autonomously scrutinise and construe video information to identify and address potential security risks. Their work examined the difficulties encountered when deploying AI systems, including the intricacy of handling substantial amounts of video data in real time, the necessity of adequate storage and retrieval techniques and the resolution of privacy and ethical issues linked with surveillance technologies. The authors emphasised the importance of continuous research and development in fully utilising image processing techniques for surveillance systems. Specifically, the focus was on identifying mobile entities in video surveillance implementations. The authors proposed a new video surveillance object detection algorithm that comprises video compression, object detection and object localisation. They also mentioned the challenges related to detecting moving objects, including but not limited to illumination variations, occlusions, complex backgrounds and camera motion. The research additionally showcased relevant studies and experimental findings to reveal the efficacy of distinct techniques for detecting mobile entities. It concluded with potential future directions and emerging trends in moving-object detection. These trends include the incorporation of deep learning (DL) techniques and the utilisation of multi-camera systems.

2.2 Old Surveillance and AI Surveillance Technologies

The global cyber threat detection system needs improvement. The cybersecurity checks of cities’ smart service applications that utilise video footage in decision-making, such as those implemented in Abu Dhabi and Dubai in the UAE, need to be considered. The challenge is that in highly crowded cities, security is mandatory to ensure normal and safe operation. Security authorities are highly interested in utilising AI to predict human behaviour because of new risks, such as cyberattacks and terrorist attacks. With AI, they can predict some crimes before they happen (Appiah et al., 2020). In 2017, malicious software attacked and disrupted the websites of the UAE’s Telecommunication Regulatory Authority (TRA), the state-owned Saudi Aramco and the Kuwaiti Ministry of Interior (Chandra et al., 2019).

The UAE’s TRA reported that the country faced 86 cyberattacks, including an attack on the customer data of the cab service company Careem, in the first quarter of 2018. The attackers leaked the data of more than 14 million customers. Given that the UAE and the Gulf Cooperation Council (GCC) region are oil- and gas-driven economies, cyberattacks on such segments can severely threaten the economy. Thus, the UAE and other GCC countries struggle to find technologies that mitigate these risks and implement effective cybersecurity strategies. According to Chandra et al. (2019), the Dubai Police reported that around one in five UAE residents was attacked by cybercriminals in 2015. Cybercrime-related reports in the country increased by 23% in the same year.

In 2016, the Kaspersky Lab ranked the UAE 8th worldwide in terms of attacks by banking Trojans. Thus, the UAE and the GCC region face an increased risk of attacks on video surveillance content. Old surveillance systems are inefficient and ineffective against new attack technologies. An overview of new and old surveillance systems is provided in Figures 2.1 and 2.2.

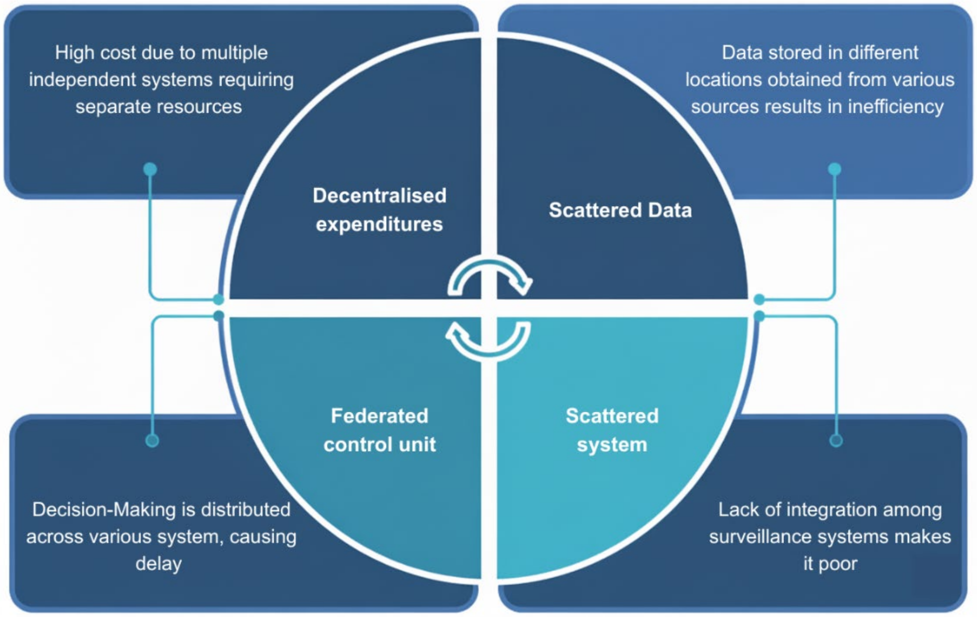

High cost due to multiple independent systems requiring separate resources
Data stored in different locations obtained from various sources results in inefficiency
Decentralised expenditures
Scattered Data
Federated control unit
Scattered system
Decision-Making is distributed across various system, causing delay
Lack of integration among surveillance systems makes it poor

Figure 2.1 Old surveillance technologies

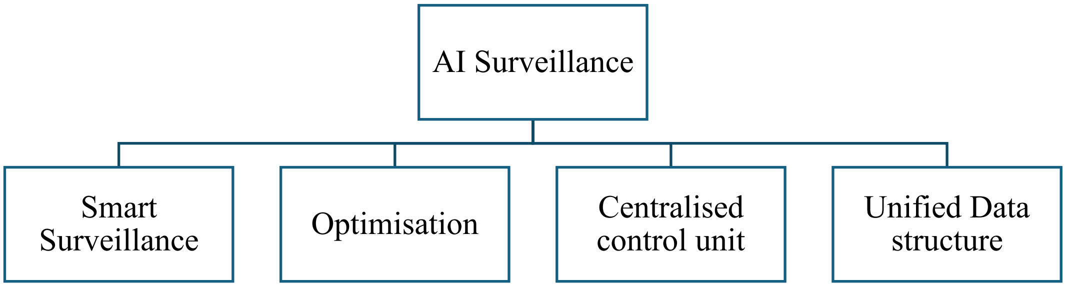


Figure 2.1 AI surveillance technologies

One of the core methodologies in this field is the application of GANs, as described in Table 2.1. The adversarial process enhances the model’s ability to generate highly realistic surveillance footage even in scenarios with sparse data. Moreover, the implementation of AI in surveillance systems enables facial recognition, behaviour analysis and anomaly detection, thereby providing comprehensive monitoring capabilities. For instance, AI can learn and adapt to normal behaviour patterns in a given environment, making it easy to identify unusual activities that may indicate security breaches or other incidents.

Table 2.1 Approaches for smart surveillance using AI with GAN

Aspects	Description	Approach
Smart Surveillance	AI enables real-time monitoring, behaviour analysis and automated threat detection, thus reducing manual oversight and errors.	GAN generates adversarial examples to test the robustness of behaviour recognition and threat detection algorithms.
Unified Data Structure	Data are centrally stored and uniformly structured, making them easy to manage, process and analyse for actionable insights.	GAN synthesises datasets to train AI models to ensure compatibility with the unified data structure and increase data diversity.
Optimisation	AI optimises resource allocation, response strategies and system performance, ensuring efficient surveillance operations.	GAN-generated scenarios can simulate edge cases to improve the optimisation of AI surveillance systems.
Centralised Control Unit	A unified control system enables fast decision-making and centralised command over all surveillance activities, increasing the overall effectiveness and coordination.	GAN can test centralised systems by generating adversarial attacks, ensuring robust centralised control and response mechanisms.

Most existing studies focused on still images and investigated other aspects of adversarial attacks and malware detection. According to Martins et al. (2020), many studies have tested different types of attacks and found that they are effective in intrusion detection. However, their practicality in attack scenarios needs to be tested (Kanim Oshi & Jacob, 2019). The study also claimed that scholars have considerably explored the defence mechanism for adversarial attacks through its efficiency in resisting them. According to a report by UAE-P (2022), the UAE ranks first in the Arab region and 13th globally in the UN E-Government Development Index (EGDI). The UAE’s overall global rank has improved by eight positions compared with the previous edition (Zawya.com, 2016) and continues to dominate in the region. EGDI measures the level of innovation and widespread adoption of electronic services in various countries. The UAE’s high ranking indicates that it has made substantial strides in utilising digital services, demonstrating its commitment to digital transformation and innovation. Such an accomplishment demonstrates the UAE’s efforts to enhance public services and improve overall effectiveness and convenience for its citizens and businesses through technology; such efforts may increase the risk of cyberattacks.

The use of AI methods in cybersecurity was examined by Jain (2021). Jain's study discussed the benefits of AI in cybersecurity, including its ability to rapidly analyse large volumes of data, identify patterns and anomalies and adapt to new and evolving threats. Moreover, it highlighted some of the challenges and limitations associated with AI in this context, including the need for high-quality training data, the interpretability of AI models and their vulnerability to adversarial attacks. The report emphasises the importance of a comprehensive strategy that integrates AI tools with human expertise in cybersecurity operations. The results support current efforts to utilise AI in creating reliable and effective cybersecurity solutions.

In the work of Zhang et al. (2020b), a variant of GAN, namely, MTGAN, was developed. MTGAN enables the generator to capture detailed information from images and produce a clear image, thereby increasing the probability of detection. The research mentioned the similarity of creating a system to increase the detection rate. However, this work differs in that it addresses the issue of adversarial GAN-based attacks on video surveillance, thereby improving the real-time GAN system. This study is unique in many ways. Firstly, studies have yet to develop an efficient GAN model that helps detect objects and attacks in a video surveillance environment. Secondly, most studies have focused on one or two aspects, such as developing GANs for adversarial attacks and detecting negative images, defending against these attacks, or simply reviewing various attack generation and detection methods. For example, the study of Zhang et al. (2019b) attempted to develop a tamper-proof system; however, the research incorporated images rather than videos. Other studies investigated the use of images for tampering detection because it is a new and emerging area of research. The literature review concludes with several research issues and gaps (Chen et al., 2020; Manjula et al., 2016).

Al-Garadi et al. (2020) highlighted the necessity of combining and securing different types of technologies. As the range of security threats and attack surfaces regarding the Internet of Things (IoT) expands, securing IoT devices and objects has become a complex issue. The research presented a thorough examination of the potential applications of ML and DL techniques in IoT security. It highlighted the benefits, drawbacks and utilisation of security for IoT. These issues were then categorised based on data, learning approaches, ML and DL for IoT security within the interrelated, interdependent, and interactive settings of IoT systems; varied security trade-offs in IoT applications; and the collaborative integration of ML and DL with blockchain technology to enhance IoT security.

Ren et al. (2020), in their recent research, also investigated adversarial attacks and defence strategies within DL. They highlighted the emergence of adversarial attacks, which leverage the vulnerabilities of DL models by introducing imperceptible perturbations to input data. The authors thoroughly examined the defence mechanisms proposed in existing literature, including adversarial training, defensive distillation, input transformation and detection-based techniques, to reduce the effect of adversarial attacks. The study delved deep into the constraints of current defence mechanisms, including their susceptibility to adaptive attacks and their applicability across various models.

Considering IoT security and attacks from secondary users, Dong et al. (2018) explored recurrent neural networks (RNNs) in practical cognitive radio network (CRN) models for primary user emulation (PUE) attacks. As shown by the model in Figure 2.3, the research highlighted that CRNs are a widely acknowledged solution, where secondary users (SUs) are permitted to share channels with licensed primary users (PUs) without causing any interference to the PUs' regular operations. However, malevolent assailants or self-interested SUs can unlawfully replicate PUs' actions to seize control of the channels.

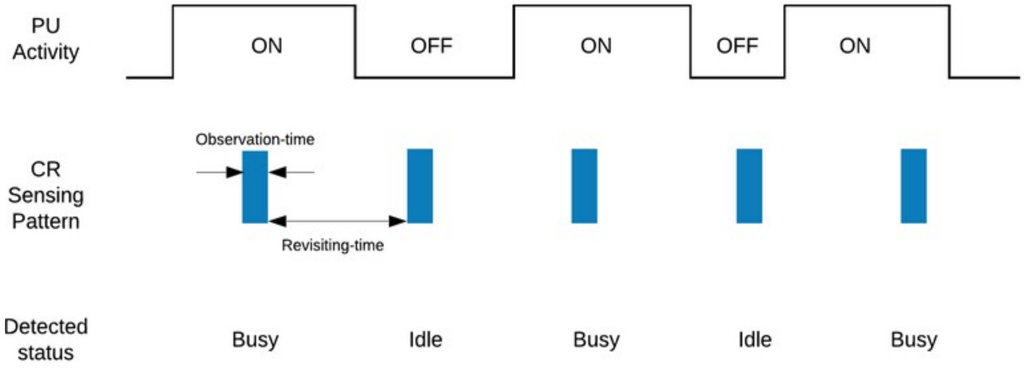


Figure 2.2 Practical CRN model of Dong et al. (2018)

The research presented how RNNs can be used to detect PUE attacks and attackers and introduced an improved version of the basic RNN algorithm, namely, the long short-term memory (LSTM) algorithm that exhibits superior efficiency in processing time series data with extended memory retention. The empirical investigation yielded a deep understanding of the distinct performance exhibited by RNNs and substantiated the efficacy of the proposed detectors.

In another survey, a prominent transportation network enterprise, Careem, operating in the Middle East region disclosed that it was subjected to a cyber intrusion that resulted in unauthorised user information retrieval. According to Alrawi (2018), the purloined information potentially encompassed individuals' names, email addresses, phone numbers and travel-related data. However, neither credit card details nor passwords were compromised. Careem, the app, assured its users that it was promptly implementing measures to address the situation and had initiated a comprehensive investigation into the incident. It recommended that its users take precautionary measures by reviewing their accounts, exercising vigilance against possible phishing attempts and promptly reporting suspicious activities. The organisation has committed to keeping its users apprised of any noteworthy developments about the occurrence.

The dataset for malicious attacks or adversarial examples is ancient and needs to be updated. A system that can detect malicious activities in real-time video surveillance and prompt action to be taken immediately needs to be created. In addition, a deep investigation is needed to strengthen generative adversarial models and develop a system that cannot be easily fooled; in this regard, a large, updated dataset will be beneficial. Many studies have highlighted the challenges posed by adversarial attacks on AI systems.

The Deepfool method (Moosavi-Dezfooli et al., 2016) is a novel technique designed to manipulate deep neural networks (DNNs) by creating adversarial examples that exhibit minimal perturbations. The iterative process of the algorithm involves the computation of the minimal perturbation required to cause the misclassification of the input sample. This computation is achieved by utilising the linearised decision boundary of the neural network. Empirical findings on standard datasets indicate that Deepfool is highly effective in producing adversarial instances that successfully mislead cutting-edge DNNs. The proponents of Deepfool also deliberated on the implications of Deepfool in evaluating the resilience of DNNs.

The Deepfool method is a simple yet effective approach for producing adversarial samples that can mislead DNNs. It exhibits remarkable accuracy in deceiving cutting-edge models. It enhances comprehension of adversarial attacks and suggests a defence mechanism in DL, providing valuable perspectives for developing resilient and safeguarded neural network architectures. However, no study has tested and evaluated attacks on generative images and videos, so future examination is required; this thesis provides such examination.

The algorithm developed by Moosavi-Dezfooli et al. (2017) is designed to calculate universal adversarial perturbations, which can mislead a diverse set of image classification models. The algorithm is optimised for a perturbation that maximises the model's prediction error across a given set of images. The algorithm updates a perturbation iteratively to minimise the L2 norm and maximise the error. This approach is effective in deceiving advanced image classification models across various datasets, including ImageNet. Empirical evidence supports the high success rates achieved by this algorithm. The authors also discussed the limitations and challenges that universal perturbations pose, including the susceptibility to image transformations and the requirement for robustness evaluation. This thesis highlighted the vulnerabilities inherent in image classification systems and enhanced the understanding of adversarial attacks and their implications in computer vision.

Carlini and Wagner (2017) focused on evaluating the resilience of neural networks in the face of adversarial assaults. The authors presented a comprehensive framework to assess the vulnerability of neural networks. This framework considers the following fundamental aspects:

(1) the minimum perturbation required to induce misclassification,

(2) the transferability of adversarial examples across multiple models and

(3) the efficacy of proposed defences in mitigating adversarial attacks.

The attack algorithm uses an optimisation procedure to identify the smallest possible perturbation that results in a targeted misclassification whilst minimising the perturbation's detectability. The study also examined the transferability of adversarial examples, demonstrating that adversarial instances generated for a particular model can potentially deceive other models. The authors mentioned the need for enhanced evaluation metrics and comprehensive analysis of defence mechanisms to propel the progress of secure and dependable neural network models.

Papernot et al. (2016) examined the vulnerabilities of DL models to adversarial attacks, which introduce minor changes to input data, resulting in incorrect predictions. Their work aimed to understand the reasons behind this susceptibility and investigate various defence mechanisms to improve the resilience of DL models against such attacks. It emphasised the need for further research and development of effective defence mechanisms, such as combining adversarial training and input preprocessing techniques, to enhance the robustness of DL models in adversarial environments. In general, the study offers valuable insights into the constraints of DL models when operating in adversarial environments. It underscores the importance of constructing robust and secure DL models to alleviate the consequences of adversarial assaults.

Engstrom et al. (2018) introduced a technique for producing adversarial examples by utilising simple alterations on authentic images. They showcased the feasibility of generating an adversarial example that can deceive CNNs into misclassifying an image by adopting a blend of rotation and translation techniques. The study assessed the susceptibility of various cutting-edge CNN structures to basic alterations. The findings revealed the efficacy of the suggested approach in producing adversarial examples that can circumvent CNNs' classification accuracy. The implications of the study's findings pertain to the security and robustness of systems based on CNNs, highlighting the necessity of developing DL models that are highly resilient and robust.

Poursaeed et al. (2018) addressed the utilisation of GANs in creating perturbations that can mislead DNNs. They suggested a theoretical structure comprising two fundamental constituents: a generator network and a classifier network. The experimental findings proved the efficacy of the proposed approach in producing adversarial perturbations that can mislead advanced DNNs. The findings are important for the resilience and security of DL models. This research can facilitate the development of additional robust countermeasures against adversarial attacks and improve the dependability of computer vision systems.

Xiao et al. (2018a) introduced a novel methodology for producing adversarial examples that exhibit resilience to spatial transformations. They utilised generative models to acquire the ability to operate a spatial transformation network that can systematically modify the input. The authors assessed the efficacy of their methodology across diverse image classification assignments and showcased its capacity to produce resilient adversarial examples that retain their adversarial characteristics amidst spatial transformations. The findings underscore the importance of incorporating spatial transformations into adversarial attack scenarios and the potential of using spatially transformed adversarial examples in assessing and improving ML models.

Furthermore, Feng et al. (2020b) introduced a novel approach for generating adversarial examples via a mask-based GAN. The approach centres on the selective alteration of a limited number of input characteristics, thereby enabling expedient and precise offensive manoeuvres. The efficacy of the approach was assessed on various benchmark datasets, and the method was juxtaposed with other contemporary attack methodologies. The findings indicated that the limited features utilised in the attack strategy yield remarkable success rates with minimal perturbations.

Cayford and Pieters (2018) investigated the efficacy of surveillance equipment. They discussed with intelligence personnel through interviews to learn about the influence and effectiveness of surveillance technology in attaining their goals. They shed light on several topics, including the efficacy of various surveillance plans, the challenges in implementing them and the moral issues raised by their application. Moreover, they examined the research’s implications and potential effect on decision-making and policy-making processes related to surveillance technologies. This report facilitates a deep understanding of the national security implications of surveillance technologies, informed conversations and decision-making processes in this area.

Sharma (2021) addressed the possible hazards and difficulties connected with widely used surveillance techniques whilst discussing the necessity for a responsible approach to digital monitoring. He suggested methods for attaining ethical and accountable surveillance applications, such as creating legislative frameworks, utilising privacy-enhancing technology and deploying accountability and transparency mechanisms. Furthermore, he stressed the importance of striking balance between national security requirements and individual privacy and civil liberties. He offered suggestions for realising a future of responsible surveillance, such as creating legal frameworks, adopting privacy-enhancing technologies and implementing transparency and accountability measures. This work promoted the creation of moral and responsible surveillance systems and added to the continuing conversation about surveillance techniques.

The global use and growth of AI surveillance systems and their consequences for personal privacy, human rights and international security were examined by Feldstein (2019). The author discussed the worldwide adoption of AI technologies and their integration into surveillance systems and argued that whilst these systems may have some advantages, they raise serious concerns about privacy invasion, attacks and power abuse. The author’s study addressed the ethical implications of AI surveillance, including mass monitoring, facial recognition technologies, data collection and large-scale analysis. It also revealed the possible effects of AI monitoring on international relations and global security and urged the development of a careful, reasonable approach to introducing these technologies. The conclusions and suggestions presented in this study contribute to the ongoing discussion on the ethical and political implications of AI monitoring.

2.3 Machine Learning and Adversarial Generation/Training

ML and DNNs are important in image classification, face recognition and object tracking and detection (Zhu et al., 2015). CNN is a standard, widely used technique for detecting faces and objects. Similar to CNN, faster region-based convolutional neural network (R-CNN) also produces state-of-the-art results (Ren et al., 2015).

Faster R-CNN is suitable for detecting objects in a large area (Kuan et al., 2017). With the successful application of DNNs, these networks can be used for object tracking and malicious behaviour classification. The system works as a unit to detect objects, as presented in Figure 2.4. The region proposal network (RPN) serves as a consideration to guide the fast RCNN module in locating objects. For instance, co-saliency in images is detected by using CNN as a basis for solving extraction problems (Alom et al., 2019).

The study conducted by Keceli (2018) examined the implementation of deep feature learning methods in identifying single and dyadic actions through viewpoint projection. The importance of comprehending the perspective from which an action is perceived is that it substantially affects the visual and kinematic properties of the action. The author suggested a viewpoint projection-based strategy for deep feature learning that utilises CNNs and RNNs to capture spatial and temporal information. The empirical findings and assessments reveal the suggested methodology’s efficacy, benefits and constraints. This work provides a valuable contribution to the progress of action recognition systems that exhibit resilience to viewpoint discrepancies.

Jan et al. (2020) conducted a thorough analysis of image retrieval methods based on the region of interest (ROI). They addressed the crucial elements and steps involved, emphasised the importance of ROI in image retrieval and assessed the benefits, drawbacks, and effectiveness of various methods. Moreover, they stressed the difficulties and unexplored areas in ROI-based image retrieval, including building reliable and scalable retrieval algorithms, managing complex scenes and accurately extracting ROIs. The study concluded by listing potential applications and future research directions.

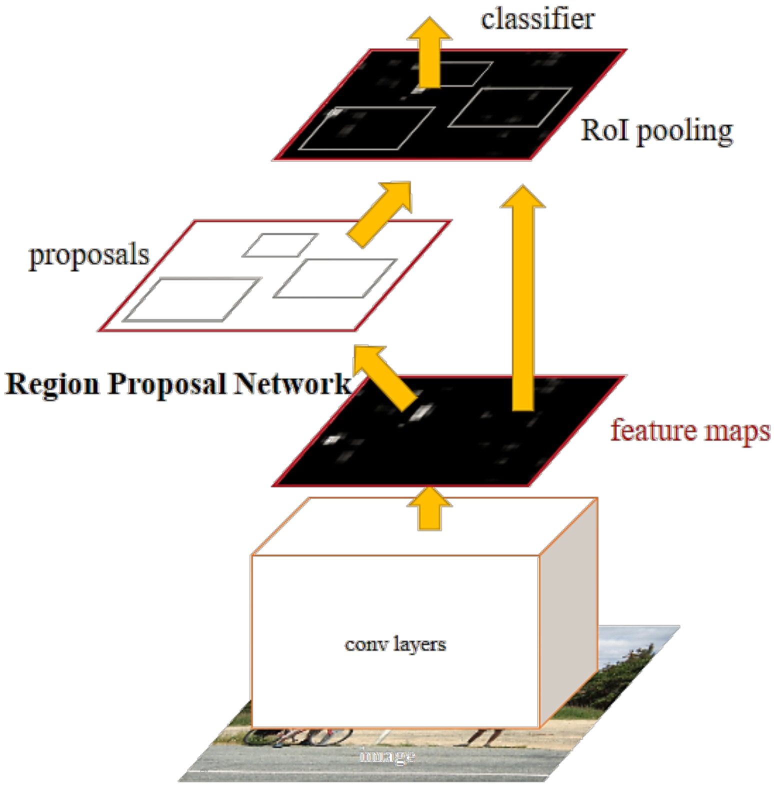


Figure 2.3 Faster R-CNN architecture (Ren et al., 2015)

2.3.1 Challenges in Image Classification/Database

Object detection in DNNs involves correct image classification. A few challenges related to the image classification necessary for database development during network training include viewpoint variation and objects that look like human occlusion. The viewpoint/post variations of an object are used to identify the object correctly from different angles/poses. Several techniques have been used in previous studies to address viewpoint variation. For example, handcrafted features were used to solve the problem, and a discriminative distance was employed to distinguish a pair of faces of the same person from another person's pair of faces by using this distance calculation, which should be smaller for the same person's face. A pose-robust technique (Li et al., 2013) using a probability elastic part (PEP) framework was applied to reduce the effect of viewpoint variations by using local binary pattern (LBP) and scale-invariant feature transform (SIFT) local appearing features. A hierarchical PEP model (Li & Hua, 2015) was utilised for fine-grained construction of face-based images or extending the previous model. DL methods produce improved results compared with handcrafted features (Chen & Chellappa, 2017).

Koraqi and Idrizi (2019) also explored the methodologies employed in detecting, recognising and monitoring objects in dynamic environments. They tackled the difficulties that arise in object analysis during motion and investigated the application of computer vision methodologies, ML models and sensor fusion techniques to effectively manage the intricacies of dynamic surroundings. This study focused on the primary phases involved in the procedure, including object detection, feature extraction, object classification or identification and motion estimation. The work also addressed the difficulties that arise in object analysis when in motion, including but not limited to occlusions, scale variations and complex background environments. The research's conclusion emphasises the possible applications and future research avenues in the domain. It recommends investigating sophisticated methodologies, including sensor fusion and DL, to enhance the performance of object detection, identification and tracking in dynamic situations. CNN produces the desired results on reducing the fine-grained classification problem because this network utilises several connected layers containing parameters integrated by kernels, which store the 3D characteristics of images. Several variations of CNNs with altered fusion strategies can be found in previous studies. Van Noord and Postma (2017) incorporated scale variance into CNN to manage pose variations by learning the deep characteristics of the image. Keceli (2018) used a pretrained CNN to map multiple views of a 3D volume and translate them to 2D features, which can be used for dyadic action recognition and distinguishing face images of the same person.

Occlusion, in addition to viewpoint variations, poses a critical challenge in detection and recognition because one object is hidden or overshadowed by the presence of another. Previously, SRC used an identity matrix as an occlusion dictionary to address occlusion, but this approach was computationally complex. Ou et al. (2014) proposed the structured model, which learns from data to reduce complexity. Wen et al. (2016) surveyed and argued that conventional holistic-based approaches are vulnerable to occlusions compared with part-/local-based approaches.

The algorithm presented by Yuan et al. (2020) integrates scale adaptability and occlusion detection for object tracking. It aims to resolve the difficulties associated with accurate object tracking in scenarios involving changes in scale and occlusions. The approach used in the algorithm effectively addresses variations in object dimensions throughout the tracking process by combining visual characteristics and flexible scale estimation. Moreover, it integrates a mechanism for identifying and managing situations of occlusion.

The algorithm was evaluated using benchmark datasets and subsequently compared with other contemporary tracking methodologies. The experiment's findings revealed that the object tracking algorithm that adapts to scale and detects occlusions exhibits superior performance compared with current techniques, particularly in situations that entail scale variations and occlusions. Its efficacy was showcased through comprehensive experimentation, underscoring its potential utilisation for diverse computer vision applications.

In the study of Ma et al. (2016), the difficulty of tracking objects in visual sequences that are affected by motion blur was addressed. The authors presented a new framework that integrates a model for appearance resistance to blur with a scheme for estimating motion blur, intending to enhance tracking accuracy. The findings indicated that the suggested technique yields a noteworthy enhancement in the efficacy of visual tracking in the presence of motion blur compared with current methodologies. This enhancement is evidenced by the method's superior accuracy and resilience in tracking objects with blurred features. The proposed method produces encouraging outcomes in enhancing tracking efficacy in motion blur, thereby contributing to the advancement of object-tracking methodologies. The vulnerability of DL algorithms to adversarial samples has led to increased interest in research on adversarial attack and defence techniques in recent years. The theoretical underpinnings, computational procedures and practical implementations of these methodologies address the unresolved issues and obstacles.

Zeng et al. (2021) examined the challenges and advancements in face recognition in the presence of facial occlusions. The authors provided a comprehensive overview of the various techniques utilised for face recognition in the presence of occlusions. Their work underscored the importance of occlusion in face recognition scenarios that occur in real-world settings. Occlusions can compromise the performance of conventional face recognition systems. The authors surveyed face recognition techniques tailored to handle occlusion. These techniques included local feature-based, holistic-based and hybrid approaches. The study also addressed the difficulties and unresolved research domains related to facial recognition in the presence of occlusion, including partial occlusions, diverse levels of occlusion and dynamic occlusion scenarios. The study also examined the potential ramifications of face recognition technology when it is confronted with occlusion in various contexts, including but not limited to surveillance, access control and forensic investigations.

Meanwhile, Ou et al. (2014) presented a resilient method for facial recognition that exploits occlusion dictionary learning and effectively tackles the difficulty of recognising faces that are partially obscured. They conducted a detailed analysis of the methodology employed in constructing an occlusion dictionary that effectively captures the diverse appearance and texture variations arising from occlusions. They showcased their empirical findings and assessments to reveal the efficacy of the suggested methodology. Moreover, the authors examined the benefits and drawbacks of the suggested methodology, including the interpretability and discriminative capability of the acquired occlusion dictionary, its scalability and efficiency and its capacity to manage various forms of occlusions. The study contributes to the progress of face recognition technologies that exhibit occlusion resilience, thereby enabling accurate identification in practical settings.

2.3.2 Object Detection Techniques

Several object techniques, such as Haar cascade, YOLO, kernelised correlation filter (KCF) and support vector machine (SVM) with histogram of oriented gradients (HOG), are discussed in this section. Phuc et al. (2019) used a Haar cascade classifier for safety equipment recognition and object detection. It is an ML algorithm trained on several positive and negative image samples via the cascade function. It consists of four stages: selection of Haar features, image creation, training with AdaBoost and classification using the cascade function.

Shetty et al. (2021) used the Haar cascade to detect the faces of registered individuals in a database from the given video input; however, the video quality, viewpoints and angles influenced the performance of the method. Cuimei et al. (2017) modified the Haar cascade to detect three classes. Ulfa and Widyantoro (2017) used the Haar cascade function for vehicle detection, and Shetty et al. (2021) found that Haar cascade function-based detection outperforms LBP and HOG.

KCF Tracking:

Zhao et al. (2021) assessed KCF for the visual object tracking approach. The study recognised the value of object tracking in several computer vision applications, including surveillance, robotics and augmented reality. The study's primary goal was to examine KCF's tracking system in depth and provide readers with a thorough knowledge of its essential elements. The authors reviewed the creation of the correlation filter, the use of kernel functions to capture the spatial and aesthetic details of the target object and the benefits and drawbacks of the KCF technique. The report presents experimental findings and analysis to confirm the KCF algorithm's efficacy and evaluate its performance compared with that of other cutting-edge tracking techniques on various benchmark datasets. The results contribute to our understanding of KCF and its application in visual object-tracking tasks.

Haar Cascade:

The Haar cascade technique was investigated by Phuc et al. (2019 for its ability to recognise safety equipment in various working settings. The authors explained how the technique works, which entails teaching a cascade classifier to detect particular patterns or objects by using positive and negative examples. To reliably identify safety equipment, they provided a system that combines preprocessing methods, feature extraction and the Haar cascade classifier. The authors presented data demonstrating the effectiveness and efficiency of the algorithm in recognising safety equipment, followed by experimental assessments to evaluate the performance of the proposed strategy. The study's conclusions had ramifications for several businesses that prioritise safety precautions and seek to use cutting-edge technologies for safety management.

Ulfa and Widyantoro (2017) focused on the pragmatic application of the Haar cascade classifier in detecting motorcycles. The Haar cascade classifier is a method based on ML that employs a series of weak classifiers arranged in a cascade to identify particular objects or patterns. The study examined the obstacles related to motorcycle detection, including discrepancies in visual characteristics, obstructions and environmental distractions. Moreover, the study suggested methodologies to overcome these obstacles and improve the accuracy of the classification system. The authors conducted experimental evaluations to validate the efficacy of the proposed approach. The results demonstrated the successful detection of motorcycles via the Haar cascade classifier. The implications of the findings are relevant to traffic monitoring systems and can enhance the safety and management of roadways.

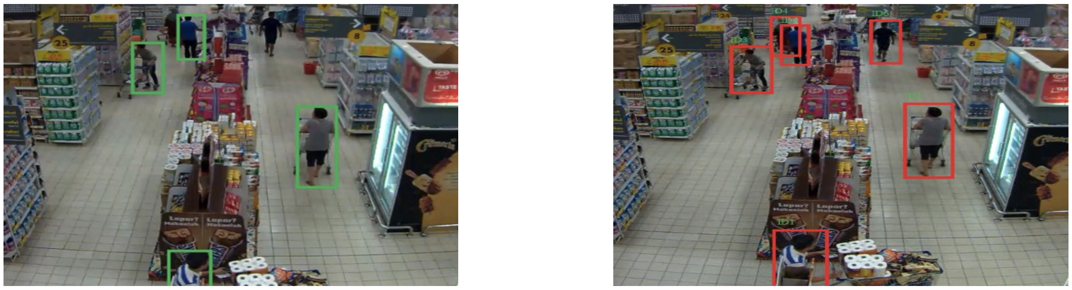
YOLO:

Compared with the Haar cascade, YOLO is a single-step approach for object detection and classification; the class is predicted after bounding box evaluation of the input (Redmon et al., 2016). Several variations of YOLO have been explored. Recently, YOLOv3 was improved by Song et al. (2021) to detect small objects efficiently and precisely by integrating Deep Sort for multi-objects. Adarsh et al. (2020) defined and described the YOLO version YOLO v3-Tiny by comparing it with existing object detection methods and recognition processes to highlight frequent algorithmic advancements and the rapid growth experienced by the field of object detection whilst aiming to enhance its speed and accuracy. This comparison showed that numerous widely used applications, such as pedestrian detection, medical imaging, robotics, self-driving cars and face detection, are instrumental in alleviating human labour in various domains.

Given the expansive scope and diverse range of cutting-edge algorithms, comprehensively addressing them in a single undertaking is challenging. Adarsh et al. (2020) provided a comprehensive overview of object detection techniques encompassing two distinct categories of object detectors. The two-stage detection approach includes R-CNN, Fast R-CNN and Faster R-CNN and prioritises accuracy. The one-stage detection approach covers YOLO v1, YOLO v2, YOLO v11 and SSD, and it prioritises speed.

C´orovic´ et al. (2018) utilised the latest YOLOv3 as an object detection technique to detect object and traffic participants. The system was designed and trained to detect five different classes of objects, including vehicles, under various weather conditions and across different driving situations. The use of advanced adversarial neural networks (ANNs) and CNNs effectively addressed the issue of slow response time when detecting objects. Meanwhile, Huang et al. (2018c) used YOLO-lite, which was designed with the help of the YOLOV2 algorithm, that is specifically designed to be used on portable devices that do not possess graphics processing units (GPUs) to detect objects with improved efficiency.

TLD is an award-winning, real-time algorithm for tracking objects in video streams (Figure 2.5). It tracks objects in a single frame, learns their appearance and detects them whenever they appear in the video. A graphical user interface system has been developed for the purpose of this research, and the results facilitate real-time tracking that typically improves over time. Zhen et al. (2020) introduced an improved methodology for visual object tracking that utilises the TLD framework. Their study suggested enhancements to the conventional TLD technique to overcome certain constraints, including the incorporation of colour-based characteristics and saliency maps during the initial bounding box estimation phase, the utilisation of adaptive appearance modelling and online classifier updating techniques at the learning-based tracking stage and the implementation of refined confidence score calculation methods at the detection refinement stage. The empirical findings indicated that the enhanced TLD algorithm attains superior tracking performance in terms of accuracy, robustness and efficiency. The proposed alterations aim to address various obstacles and constraints encountered by the conventional TLD algorithm, thereby improving tracking accuracy and resilience. The empirical assessments validated the efficacy of the suggested algorithm, establishing it as a prospective method for visual object tracking across diverse domains.



(A) 4/6 people detected YOLO only (B) 6/6 people detected by
YOLO with TLD

Figure 2.4 (A) Illustrates the "Basic" YOLO performance, likely showing a standard detection with potential issues like a less accurate bounding box or a missing object ID., and (B) Illustrates the "Improved" YOLO performance, which is enhanced by the TLD (Tracking-Learning-Detection) algorithm.

Shin et al. (2020) argued that deep feature approaches are ineffective for real-time object tracking and thus used KCF. KCF uses kernel tricks in combination with circulant matrices to reduce the computational complexity of object detection and tracking, and it produces good results (Yadav & Payandeh, 2018).

TLD Algorithm:

In the work of Abdali et al. (2017), the utilisation of the TLD algorithm on a field-programmable gate array (FPGA) was examined to achieve hardware acceleration. The authors emphasised the potential of FPGA-based acceleration to enhance the effectiveness and efficiency of the TLD algorithm. The study highlighted the importance of real-time object tracking in various domains, including surveillance and robotics, as well as the computational challenges it entails. The authors presented their empirical findings and assessments of the FPGA-accelerated TLD algorithm. The performance of the algorithm was compared with that of software-based implementations. The study also examined the advantages and challenges associated with FPGA-driven acceleration for the TLD algorithm, highlighting the potential for real-time tracking applications, reduced energy consumption and improved scalability. The research concluded by showcasing the utilisation of FPGA technology for hardware acceleration of the

TLD algorithm. The study provided valuable insights into the potential of FPGA for executing algorithms with high efficiency and performance.

Han et al. (2018) conducted a thorough investigation of contemporary DL methodologies utilised for identifying prominent and category-specific objects. They assessed notable models, including Faster R-CNN, YOLO and SSD. They emphasised the models' respective merits, drawbacks and efficacy in terms of detection accuracy and velocity and investigated sophisticated methodologies for tackling specific object detection tasks, such as detecting salient objects and objects belonging to particular categories. The research also highlighted the present obstacles and forthcoming prospects in the domain, including the assimilation of multi-modal data, instantaneous execution and comprehensibility of DL models. This work is valuable for scholars and professionals engaged in the domain of object detection through DL methodologies.

Feature Augmentation and Sampling Adaptation Algorithm:

Yildirim and Süsstrunk (2015) introduced a technique to achieve efficient and accurate identification of prominent entities in images whilst accounting for their dimensions. The feature augmentation and sampling adaptation (FASA) algorithm under consideration prioritises three fundamental elements, namely, speed, accuracy and size consciousness. The study explained the algorithm in detail, including the computational procedures and feature representation employed. The authors' approach was compared with notable object detection methods on benchmark datasets. The authors discussed the algorithm's practical applications, including image compression, object recognition and content-based image retrieval. The results contribute to the progress of salient object detection methodologies and offer valuable perspectives for the enhancement of computer vision systems that are efficient and accurate. To attain high accuracy and resilience, Girshick et al. (2014) developed a DL strategy for object recognition and semantic segmentation tasks. An RPN, a classifier and several convolutional and pooling layers make up the proposed model. The authors conducted extensive tests on benchmark datasets, such as PASCAL VOC and MS COCO, to show the viability of their strategy. The influence of several model elements, including the region proposal procedure and the usage of pre-training and fine-tuning, was also discussed. The results substantially influenced subsequent studies and paved the way for development in these fields.

Hosang et al. (2015) considered the elements that affect how well detection suggestions perform in computer vision tasks. They examined various detection proposal criteria, including quality, diversity and geographic coverage, to determine which ones contribute to good object detection performance. The research presented an assessment framework that enables a thorough examination of several cutting-edge detection proposal algorithms. The results showed that appropriate balance amongst high recall, minimal overlap and accurate localisation is necessary for effective detection approaches. The authors also examined trade-offs between accuracy and speed when analysing the computing effectiveness of the proposed methods. The results enhanced object detection performance and inspired the creation of subsequent detection proposal approaches.

R-CNN:

Faster R-CNN, a trainable end-to-end framework that combines an RPN with a CNN for object identification, was proposed by Ren et al. (2015). The performance of Faster R-CNN on benchmark datasets was assessed, and the study provided comprehensive descriptions of the architecture and training processes for RPN and CNN. The outcomes showed that Faster R-CNN maintains real-time processing rates whilst achieving state-of-the-art accuracy. The study also addressed methods to enhance the accuracy and effectiveness of the system and offered insights into how various design decisions and hyperparameters affect the performance of Faster R-CNN. The Faster R-CNN architecture has inspired subsequent developments in object identification and has become a critical part of it.

Region Average Pooling Approach:

By adding contextual information, Kuan et al. (2017) developed region average pooling (RAP), a unique method for enhancing object detection. RAP collects contextual information that can help with accurate object recognition by aggregating data from the target region and its nearby regions. The RAP algorithm’s capability to effectively incorporate contextual signals into current object detection frameworks was highlighted by the authors as they outlined the RAP algorithm’s architecture and implementation details.

The authors conducted trials on benchmark datasets, such as PASCAL VOC and MS COCO, to gauge RAP’s efficacy. The outcomes showed that adding contextual data via RAP enhances the functionality of object detection algorithms, resulting in high localisation and accuracy. The research compared the outcomes with those of other cutting-edge object identification techniques and explored the benefits and drawbacks of the suggested strategy. RAP’s possible expansions and future research directions were also highlighted.

YOLO-SORT:

Bathija and Sharma (2019) provided an innovative framework for visual object detection and tracking. The framework integrates two widely used techniques, namely, YOLO for object detection and SORT for object tracking. The study showcased its empirical findings on standard datasets to assess the efficacy of the suggested framework. It used various measures, such as tracking accuracy, computational velocity and resilience to obstructions and alterations in object manifestation, to measure performance. The findings indicated that the YOLO-SORT framework attains a comparable level of performance in terms of detection accuracy and tracking dependability. The aforementioned methodology has promising prospects for utilisation in various fields of computer vision, such as autonomous vehicles, surveillance systems and other related domains.

YOLO–Attentive Capsule Network:

The YOLO–attentive capsule network (YOLO-ACN) was introduced by Li et al. (2020b) as a new methodology for resolving the difficulties associated with identifying small targets and obscured objects in computer vision applications. YOLO is known for its ability to detect objects in real time. By contrast, the ACN module aims to improve the detection accuracy for objects that are small or partially obstructed. The study presented a thorough assessment of YOLO-ACN's ability to detect small targets and occluded objects on established datasets. The empirical findings indicated that YOLO-ACN surpasses current leading approaches in accuracy and resilience when faced with difficult situations. It is appropriate for implementation in various domains, including surveillance, autonomous vehicular navigation and object trajectory monitoring.

Ryu and Chung (2021) presented an entirely new detection model that aims to address the issue of identifying occluded objects by utilising the YOLO algorithm. The detection accuracy in scenarios where objects are partially obscured is improved by the model through the use of hard-example mining and augmentation policy optimisation techniques. The study thoroughly assessed the proposed framework on standard datasets for detecting occluded objects. The empirical findings indicated that the proposed model exhibits superior performance compared with the conventional YOLO algorithm and other advanced techniques in terms of accuracy and resilience under occlusion circumstances. The utilisation of hard-example mining and augmentation policy optimisation methodologies substantially enhances the accuracy of occluded object detection in the model.

Multiple Object Tracking Algorithms:

Park et al. (2021) conducted comprehensive examination and evaluation of diverse DL methodologies and applied these methods to the context of multiple object tracking (MOT) assignment. Their study classified DL-driven methods for MOT into distinct categories in accordance with their fundamental constituents, including detection models, feature representations, association algorithms and online tracking strategies. The study employed benchmark datasets that are frequently utilised for the assessment of MOT algorithms. Furthermore, a comparative analysis of the efficacy of various DL techniques was conducted on these datasets. The study culminated by providing a synopsis of the patterns and unresolved issues in DL-driven MOT, identifying the domains that necessitate additional investigation and proposing prospective avenues for forthcoming advancements.

Altered R-CNN:

The framework presented by Khan et al. (2019) integrates the Faster R-CNN object detection algorithm with deep appearance features, resulting in a multi-person tracking system that is both accurate and resilient. The Faster R-CNN algorithm was applied to detect individuals, and deep appearance features were extracted from the detected bounding boxes by using a pre-trained DL model. The authors conducted experiments on benchmark datasets to assess the efficacy of the proposed approach and compared its outcomes with those of other state-of-the-art tracking methods. The experiment’s findings indicated that integrating Faster R-CNN and deep appearance features leads to enhanced tracking performance, resulting in increased accuracy and resilience in tasks involving the tracking of multiple individuals. The study delineated the benefits of the proposed methodology and underscored its prospective implementations in diverse fields, such as monitoring, human–machine communication and conduct evaluation.

Noise-Modulated GAN:

The noise-modulated GAN (NM-GAN) methodology was introduced by Chen et al. (2021) to identify anomalies in video data. The NM-GAN algorithm surpasses current leading-edge techniques in anomaly detection. Ruiqiang et al. (2021) presented an optimisation framework that utilises the capabilities of GANs to enhance the identification of diminutive entities. Empirical findings indicated that the optimisation framework yields a notable enhancement in the detection of diminutive entities. Cheng et al. (2020) proposed a tracking framework that utilises an adversarial learning network to acquire attention mechanisms for accurate and resilient object tracking.

GAN-based R-CNN:

Huang et al. (2018c) proposed a two-phase methodology that combines GANs and Faster R-CNN to achieve reliable and accurate traffic sign detection. The module utilises GANs for data augmentation and Faster R-CNN for detection. This study introduced a novel methodology for detecting traffic signs by integrating GAN-based data augmentation with the Faster R-CNN algorithm. This integration improves the detection accuracy and resilience of the system, making it suitable for practical implementation in intelligent transportation systems. Prakash and Karam (2021) conducted empirical assessments on standard datasets and demonstrated the efficacy of the proposed methodology, which surpasses conventional object detection techniques on images characterised by limited resolution, noise and compression distortions.

Feature- or Marker-Based Tracking:

Feng et al. (2018) presented a technique to address the co-occurrence of scale variation and occlusion in the context of object tracking. The methodology utilises a fusion of visual characteristics and tracking algorithms to enhance the resilience and accuracy of the tracking mechanism. The feature-based representation of the tracked object, which encompasses the encoding of appearance and scale information, was discussed. Furthermore, the study introduced an innovative algorithm that integrates contextual information to address occlusion scenarios. The empirical findings indicated that the suggested approach performs satisfactorily in demanding tracking scenarios that entail scale variation and occlusion. The results demonstrated enhanced tracking accuracy and resilience compared with existing methods. The study's conclusion highlighted the importance of addressing the challenge of dealing with scale variation and occlusion in visual object tracking. The authors also proposed potential avenues for future research to advance object-tracking capabilities.

2.3.3 Object Tracking/People Tracking

In AI analytics and video processing, people tracking is achieved by detecting and tracking people in a region of interest or within a defined zone. The region of interest needs to be defined at the first stage, followed by the detection of a person in that region of interest and the assignment of a unique ID, which is counted and tracked.

TLD:

A new entry in that region of interest is dealt with in a similar process. In the event of the exit of one of the persons or entries, that person is no longer tracked or counted by ID. Kalal et al. (2010) proposed TLD, which integrates three tasks, namely, tracking, detecting and learning, such that the object is tracked frame to frame, obtaining localisation information and appearance attributes and modifying the tracker when needed. The learning calculates the error of the detector and updates it to reduce the sum of the errors. Velastin et al. (2020) argued that the TLD proposed by Kalal et al. (2010) is an unsatisfactory tracking method and that the struck technique proposed by Hare et al. (2015) is less accurate than other techniques. Rajjak and Kureshi (2019) conducted a comprehensive survey of recent advancements in object detection and tracking techniques designed to cater to the demands of high-resolution video applications. The authors discussed the progress of contemporary algorithms for object detection and tracking, including models based on DL, methods founded on features and approaches that combine both. The authors also emphasised the fundamental constituents and phases involved in object detection and tracking systems designed for high-resolution video, including the generation of object proposals, feature extraction and motion estimation. The study culminated in delineating prevailing research patterns and forthcoming avenues for research in the domain. These avenues include amalgamating multi-modal data, devising adaptable algorithms and investigating edge computing for instantaneous handling of high-resolution video.

DNNs are prominent in many fields of computer vision and image/video processing nowadays, and CNN and GAN are praised for their flexibility, ability to train large datasets and improved performance. In GAN, a con artist (based on CNN) fights against the officer (another CNN) to reach a new standard of realism. CNN is very good at detecting small and large objects in still images. Meanwhile, adversarial networks generate scalable features selectively, so GAN is ideal for real-time environments (Murthy et al., 2020).

***CNN*:**

The resilient object detection algorithm presented by Vorobjov et al. (2018) is tailored for high-resolution videos. The researchers proposed a CNN-based methodology to augment the accuracy of object detection in high-resolution video frames. The algorithm employs a customised CNN structure for object detection in high-resolution videos. The network is trained to acquire distinctive features that facilitate the accurate detection and localisation of objects. The study empirically evaluated the efficacy of the proposed algorithm and tested its efficiency against other contemporary techniques for object detection by using high-resolution video datasets. The authors discussed the algorithm’s practical applications in various domains, including surveillance, video analysis and autonomous systems. The study's results contribute to the advancement of object detection methodologies and provide valuable insights for the development of resilient computer vision systems in high-resolution video contexts.

CNN with Extreme Learning Machine:

Another research by Alshalali and Josyula (2018) involved transfer learning as a modern method that utilises pre-trained CNNs as an application to recognise the object and classify the image. With advancements, different techniques of fine-tuning pre-existing CNNs and the optimal timing to implement these techniques have become a popular subject for researchers. However, Alshalali and Josyula (2018) stated that the utilisation of transfer learning is influenced by how it affects every technique in terms of training time and test set accuracy; they also assessed the efficacy of extreme learning machine (ELM) in comparison with connected layers with respect to training time and testing accuracy in the context of transfer learning.

GANs:

GANs are beneficial in selecting and enhancing the resolution of images with small objects. The generator of GAN improves the resolution of small objects to a detailed one that is nearly indistinguishable from a real object, thereby deceiving the discriminator. This approach can be used to aid and strengthen the objective of creating a tamper-proof system. Further research in the field of AI and DL is needed to investigate other techniques that can be used to develop cost-effective, efficient methods, such as those with high detection accuracy and low computation or training time (Kalirajan & Sudha, 2015a).

Segmentation:

Owing to technological advancements, neural networking now plays a crucial role in the development of the medical field. According to Kumar et al. (2017), nuclear segmentation in digital microscopic tissue images has the potential to facilitate the extraction of superior-quality features for nuclear morphometrics and other analytical purposes. Segmentation using ML techniques can be applied to diverse nuclear morphologies; however, it necessitates a collection of image datasets. This study presented a publicly available dataset of hematoxylin and eosin-stained tissue images containing over 21,000 annotated nuclear boundaries. It also introduced a novel metric for assessing nuclear segmentation outcomes. The proposed metric was designed to cohesively consider object- and pixel-level inaccuracies. Moreover, a DL-based segmentation approach that prioritises the detection of nuclear boundaries was presented.

Jia and Hou (2021) also highlighted the importance of technological advancements in defining and identifying objects/traits. Their study presented a novel algorithm based on DL techniques for assessing physical health and fitness levels in adolescents. The collected data included age, body fat, body mass index, lean body mass, years to peak height velocity, flexibility, jumping ability, aerobic fitness level and sprinting performance. The efficacy of the system was assessed through empirical testing and analytical scrutiny.

FPGAs and Systems on Chip:

A thorough analysis of research that investigated the identification of objects in photos and videos by using various DL approaches and embedded platforms was conducted by Murthy et al. (2020). The analysis explored different DL models, including Faster R-CNN, YOLO and SSD, and evaluated their advantages and disadvantages in object identification tasks. It also examined how DL models can be integrated with embedded platforms, such as FPGAs and system-on-chip gadgets. The study examined several experiments and offered information on how various DL models and embedded platforms perform in object detection tasks. Real-time object identification, edge computing and resource-constrained situations were some of the issues and future objectives covered in this study. The study provided insights into cutting-edge methods and potential paths for further research, serving as a valuable resource for academics and industry professionals working in the field of object detection. The study conducted by Yu et al. (2018) focused on the advancement of DL models that employ generative techniques for video prediction. The model under consideration operates by processing a series of input frames and producing subsequent frames through the assimilation of the fundamental dynamics and patterns present in the video data. The experiment results proved the efficacy of the suggested methodology in producing authentic forthcoming video frames. The study also examined the constraints and potential developments of deep generative video prediction, including extended prognostications, managing obstructions and intricate environments and simulating supplementary limitations or priors. The proposed approach produces encouraging outcomes and presents opportunities for additional investigation in the field of video prediction and generative modelling.

Kim and Jung (2017) presented a novel methodology for the swift identification of individuals through the fusion of background subtraction and DNNs. The method combines the advantages of background subtraction and DNNs to effectively partition foreground regions in video frames. DNNs are utilised to classify the partitioned regions as either persons or non-persons. The study empirically assessed the efficacy of the hybrid framework by using parameters, such as detection accuracy, computational speed and resilience to fluctuations in illumination and contextual factors. The authors also examined the pragmatic applications of their framework in real-world situations, such as video surveillance systems and intelligent urban applications. This study contributes to the advancement of person detection methodologies and provide valuable insights for the development of reliable, adaptable systems in various fields that rely on person detection in the context of big data.

Kanimozhi and Jacob (2019) devised an intrusion detection system that leverages AI techniques and cloud computing to enhance the accuracy and efficiency of network intrusion detection on the CSE-CIC-IDS2018 realistic cyber dataset. The proposed methodology utilises AI, specifically ML algorithms, to analyse network traffic data and detect plausible intrusions. Hyperparameter optimisation tuning was employed to improve the efficacy of ML models, and cloud computing was leveraged to process the extensive CSE-CIC-IDS2018 dataset in a scalable and efficient manner. The study presented a comprehensive account of the experimental configuration and methodology employed to assess the efficacy of the suggested intrusion detection mechanism. The authors compared their approach with other established intrusion detection techniques by utilising metrics, such as detection accuracy, false positive rate and computational efficiency. The authors examined the advantages of employing cloud computing for network intrusion detection; such advantages include the capability to manage substantial quantities of network traffic data, scalability and cost efficiency. The results of this study contribute to the area of network security and offer valuable perspectives for the creation of resilient and adaptable intrusion detection systems through the utilisation of AI methodologies and cloud computing assets.

YOLO with Monocular Depth Estimation:

A new methodology that integrates the YOLO algorithm with monocular depth estimation (MDE) was proposed by Yu and Choi (2021) to improve the accuracy and spatial comprehension of object detection assignments. The importance of object detection in diverse computer vision domains, including autonomous driving, robotics and surveillance systems, was underscored in the study.

The proposed YOLO MDE methodology integrates MDE into the YOLO framework to enhance the accuracy of object detection and classification within an image. The authors discussed the utilisation of CNNs for object detection and MDE tasks. They experimentally evaluated the performance of their approach on benchmark datasets. The authors also addressed the practical implications of YOLO MDE. They highlighted its importance in scenarios involving autonomous driving, where accurate depth estimation plays a crucial role in understanding the surroundings and ensuring safe navigation. This work contributes to the advancement of object detection methodologies and offers valuable insights for enhancing computer vision systems.

Zhihuan et al. (2018) designed a methodology that employs the YOLO model to achieve effective target detection in high-resolution remote sensing images. The proposed approach utilises the YOLO model, a cutting-edge algorithm for object detection, to identify targets present in images obtained through remote sensing. The study presented a thorough account of the target detection framework based on YOLO, including its network structure and training methodology. The authors examined the pre-processing procedures and the identification of suitable anchor boxes for managing diverse target dimensions in remote sensing imagery. Moreover, they presented empirical evidence that demonstrates the efficacy and efficiency of their methodology on established datasets. Expeditious and accurate target identification is pivotal in enhancing the efficiency and efficacy of remote sensing applications.

***YOLO*:**

The YOLO algorithm was introduced by Redmon et al. (2016) as a means of achieving high accuracy and real-time performance in object detection tasks through a unified approach. The importance of real-time object detection in diverse computer vision domains, including autonomous driving, surveillance systems and robotics, was emphasised in the study. The YOLO algorithm partitions the input image into grids and subsequently generates predictions for bounding boxes and class probabilities by directly analysing the grid cells. The paper presents an elaborate account of the YOLO algorithm, including the network structure and training methodology. The trade-off between accuracy and speed was analysed by the authors, along with the strategies implemented to address variations in object scale and aspect ratio. The practical implications of real-time object detection facilitated by the YOLO algorithm, including its application in real-time video analysis, were also deliberated by the authors. The YOLO algorithm's ability to provide high accuracy and rapid processing is instrumental in the advancement of real-time computer vision applications.

Microsoft COCO Dataset:

The Microsoft COCO dataset was introduced by Lin et al. (2014) as a comprehensive image dataset that caters to the requirements of object detection, segmentation and captioning tasks. It is a large-scale dataset specifically designed for these purposes. The dataset comprises more than 200,000 images and 80 distinct object categories. It is accompanied with meticulous annotations for object bounding boxes, object segmentation masks and image captions, all of which are of superior quality. The study outlined the development of the COCO dataset, including the methodology for image acquisition, annotation procedures and quality control measures. The authors also provided benchmark results obtained using the COCO dataset, showcasing its efficacy in assessing algorithms for object detection, segmentation and captioning. The COCO dataset has gained substantial popularity in the computer vision community since its inception and is now widely employed as a standard reference for assessing and contrasting diverse algorithms.

***DNN*:**

Cao et al. (2018) claimed to solve the problem of computational complexity by proposing a novel framework. The accuracy rate they obtained was 0.6, which is unsatisfactory. Developing an ML system is a complicated yet effective task that should be handled carefully, and certain measures should be taken (Barreno et al., 2006). Meanwhile, the application of DNNs in human behaviour recognition and taking actions, such as self-driving, has raised questions concerning safety and trustworthiness (Huang et al., 2020b). DNNs may make the wrong prediction in the presence of adversarial images. The problem of false positives and false accuracy in detection must be addressed when developing a GAN (Pan et al., 2019b).

The most popular algorithms used to generate adversarial images or instances are as follows:

A. *Fast gradient sign method (FGSM):*

It is a single-step simple method where the adversarial noise or perturbation is calculated by analysing the gradient's direction (T. Huang et al., 2020). The generation of adversarial examples is formalized in Equation 2.1, which encapsulates the Fast Gradient Sign Method (FGSM) framework.

$$X_{adv} = x + \in . sign(\nabla x J(\vartheta, x, y) \qquad (2.1)$$

Where*:*

- X_{adv} : Adversarial example
- x: Original input image
- y: True label for x
- $\in$: A small scalar that constraints the perturbation size
- ϑ: Parameters of the model
- *J* (ϑ, x, y): The cost function used to train the model
- ∇x: Gradient of the cost function with respect to the input x
- sign (·): The sign function, which returns ±1 for each element of the input

Every input image is given a label noted by y in Equation (2.1) for its correct classification. The noise in the image aims to attack the system to misclassify the image. The noise is kept as small as possible, which is performed by using

$$\|x - x_{\text{adv}}\|_\infty < \epsilon$$

B. L–Broyden–Flecther–Goldfarb–Shanno Method:

In the L–Broyden–Flecther–Goldfarb–Shanno (L-BFGS) method, an adversarial attack on an image is formulated as a constrained optimization problem X. Li & F. Li (2017)

$$\min_{\delta} \; c \cdot \|\delta\|_2 + J(x + \delta, t) \qquad (2.2)$$

$$\textit{Subject to} \qquad x + \delta \in [L, U]^m \qquad (2.3)$$

Where:

- δ: The adversarial perturbation to be found
- $J(x + \delta, t)$: The loss function for the adversarial example $x + \delta$ with target label t
- $\|\delta\|_2$: The L2-norm of the perturbation
- $c > 0$: A constant trading off the perturbation magnitude and the loss, often found via line-search
- $[L,U]^m$: The valid range (e.g., [0, 255]) for each of the m pixels in the image

C. Carlini–Wagner Method:

Equation 2.4 formulates the minimisation problem consisting of a squared Euclidean distance term and a weighted regularisation term applied to the transformed parameters. The Carlini–Wagner (CW) attack employs three distance metrics (L_0, L_2 and L_∞). CW with input image *x*, target *t* and adversarial attack L_2 is introduced as follows:

$$min \left\| \frac{1}{2}(\tanh(\omega) + 1 - x \quad ||\binom{2}{2} + c.f\left(\frac{1}{2}(\tanh(\omega) + 1)\right), \qquad (2.4)$$

$$(x^{adv}) = max\left(max\{Z\left(x_i^{adv}\right): i \neq t\} - Z(x^{adv})_t, -k\right). \qquad (2.5)$$

In Equation 2.5, the objective function is given by *f. Z (.)*denotes the log value, and *w* is the optimisation variable. The confidence parameter is given by *k* (Massoli et al., 2021).

Another version, such as L∞, involves an adversarial perturbation variable, which is given by

$$minc\, f(x+\delta)+\sum_i\ [(\delta i-\tau)^+], \qquad (2.6)$$

where threshold level τ monitors the adversarial noise or perturbation introduced to the input image to train the network on adversaries. The above-mentioned methods are used to generate and train the model on adversarial samples. Many studies have focused on these popular methods of adversarial generation to implement effective adversarial detection techniques.

2.3.4 Adversarial Detection Approach

Massoli et al. (2021) adopted a detection approach that uses classifier *f θ(.): X → C*, where C refers to the labels allowed for input *X*, such that *X ⊆ Rd* denotes the dimension of the input.

$$f\ \theta(x)\ =\ f^n(0_{n-1};\ \theta_n)\ \circ\ f^{n-1}\big(0_{n-2;\theta^n}\ 1\big)\circ\ \ldots\ \circ\ f^{\circ}(x;\theta_0), \qquad (2.7)$$

where x is the input of the classifier and *oi* denotes the output value for *fi*, such as the i^{th} layer having parameters denoted by *θi*. An overview of the proposed detection technique is given in Figure 2.6. The basic model is shown on the left, and a single-embedding process is given on the right.

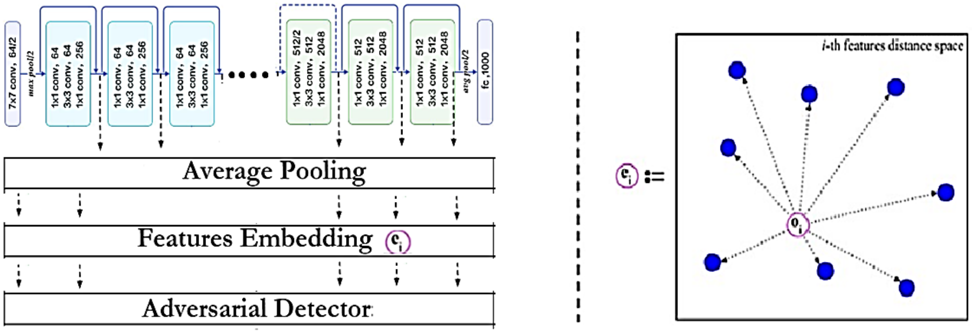


Figure 2.5 Detection model (above) and embedding process (down) (Massoli et al., 2021)

Vector *e_i* is shown in terms of its distance from the pivots, medoids and centroids. Pivots are the coordinates representing the reference points distanced from the input. They are evaluated using only the training dataset with two options or classes, namely, centroids and medoids, which are given by Equations 2.8 and 2.9, respectively.

$$p_i^j = c_i^j = \frac{1}{|B_c|} \cdot \sum\nolimits_n^{|B_c|} o_{i,n}^j \qquad (2.8)$$

$$p_{i_j} = m_{i_j} = \arg\min_c \sum_{n=1}^{|B|} ||\, o_n - o_{i_j}||_2 \qquad (2.9)$$

For class *j*, in the *ith* layer, *o(i,n)j* indicates the output of class *j* in layer *i* or the *ith* layer, where |*Bc*| denotes the cardinality of the class. The study incorporated the CW method, basic iterative method and MI-FGSM to generate an adversarial attack. The maximum adversarial noise or perturbation is denoted by ϵ, that is, ϵ = 0.03, 0.07, 0.1,0.3 (Ozbulak et al., 2020). The value of noise or perturbation can be as low as 0.07 and as high as 0.3. It is kept at less than 1. The adversarial images are generated using the mentioned algorithms, and the process is illustrated in Figure 2.7.

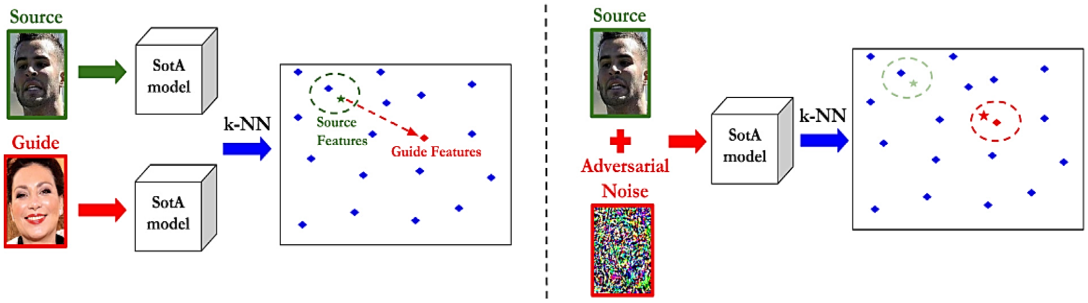


Figure 2.6 Generating adversarial samples (Massoli et al., 2021)

The authors used a *k* nearest neighbour (kNN) classifier to view the problem of generating an adversary as an optimisation issue and to guide the source, as shown in Figures 2.7 and 2.8. The left side of the figures shows the source and guide before the attacks, and the right side presents the results after adversarial noise has been added. SotA, a state-of-the-art feature extraction model, was used to extract features from the input image. Threshold values were applied during adversary generation. The adversarial noise limits or threshold values were set between 5 and 10 for this case, that is, δ ∈ (5, 7) and (10, 5).

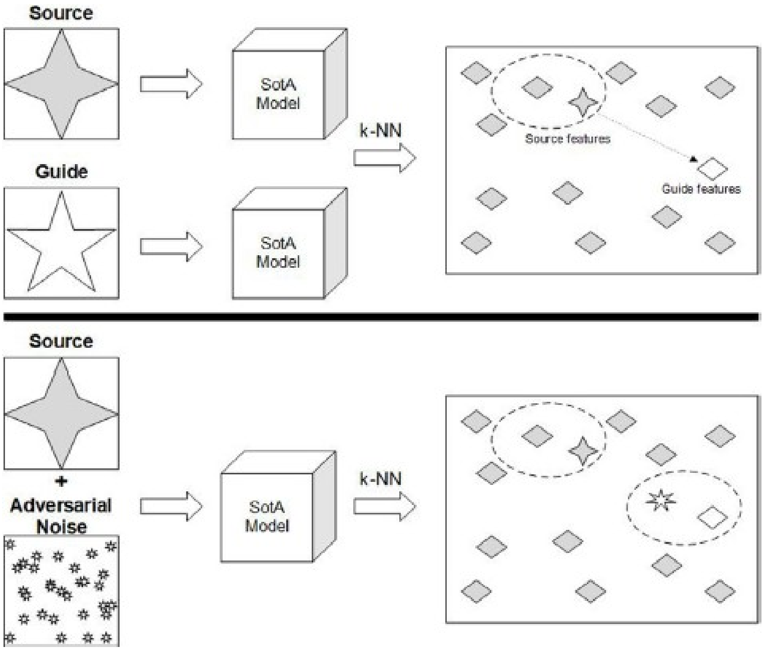


Figure 2.7 Generating adversarial samples

Figure 2.9 shows that the adversarial images look similar to the source or input image. The highest accuracy with a maximum perturbation of 10 was 96.3% for supervised training and 96.8% for unsupervised training.

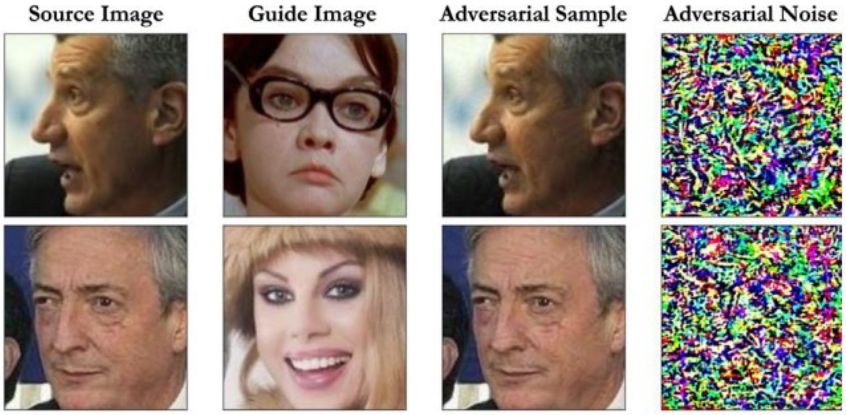


Figure 2.8 Adversarial samples with two thresholds (Massoli et al., 2021)

Table 2.2 shows a comparison of various adversarial attacks and the kind of adversarial attacks or other threats to fool the network. The training or the core method used to generate the threat is discussed, followed by the nature of the attack, such as targeted, untargeted, or both. The information required to be accessed by the generated adversarial attacks, such as either model parameter information or logits, which imply inputs to the *softmax* layer of the model, is also highlighted. The distance metric used for each adversarial generation method is identified, followed by the vulnerability of the models, which are found to be easily fooled by the respective adversarial attacks or threats.

Table 2.2 Comparison of different adversarial attacks and vulnerable models

Research/ Study	**Attacks/Threats**	**Training/ Core Method**	**Targeted/ Untargeted**	**Accessed Information**	**Distance Metric**	**Vulnerable Models**
Moosavi-Dezfooli et al., 2016	DeepFool/ perturbations to fool the network	Iterative linearisation	Both	Model parameters	Lp, $p \in [1, \infty)$	Deep neural networks
Moosavi-Dezfooli et al., 2017	Universal adversarial perturbations	Generalising DeepFool to create universal adversarial attacks	Both	Logits	L2 (universal perturbation)	Deep neural network classifier
Carlini & Wagner, 2017	High-confidence adversarial examples, targeted misclassification	Adam optimiser	Both	Logits	L0, L2, L∞	Distilled/undistilled neural networks
Papernot et. al., 2016	JSMA, adversarial perturbations	Jacobian saliency	Both	Model parameters	L0	Deep neural networks
Engstrom et al., 2018	Misclassification by rotation and /or translation	Natural transformations	Both	Logits	n/a	CNN
Poursaeed al., 2018	Image dependent and universal perturbations	GAN-based adversarial generation	Both	Logits	Lp (universal perturbation)	FCN
Xiao et. al., 2018b	Spatially transformed adversarial	Minimising adversarial and Lflow loss	Both	Logits	$L_{flow}(\cdot)$ (measuring	Deep neural networks

	examples; high-quality, sophisticated adversarial attacks				geometric distortion)	
Feng et al., 2020a	Few-feature-attack-GAN (FFA-GAN), black-box attack	Mask mechanism, GAN training	Targeted	Logits	L0, L1	Machine learning models

2.4 Generative Adversarial Attacks

Neural networks are everywhere, and hackers can hack them easily. Adversarial attacks on neural networks involve strategically developed noise that is created to fool the network. With regard to a classification algorithm with decision-making boundaries, the network is corrupted by the introduction of noise. Figure 2.10 illustrates instances of adversarial attacks. The green dots represent data points that have been trained and classified by the system. The system is trained using parameters and weights or simple characteristics, based on which a decision boundary is formed, as shown in Figure 2.10. The orange data points represent adversaries that have misclassified decisions because of perturbations and attacks that occurred in the system. Recently, ML approaches have been applied in different areas of the world; their application has created risks of adversarial attacks.

A study used NSL-KDD datasets and obtained effective results by employing a GAN framework called IDSGAN responsible for generating the adversarial nature of attacks and deceiving IDS (Lin et al., 2022). The first attack on surveillance video was done by changing the pixel values of the image to fool the classifier into showing the wrong class. Other attack approaches included patches that were applied to the object to fool the classifiers and detectors. Security systems are vulnerable because adversarial attacks cannot be detected by surveillance cameras. Approaches based on ML depend on the identification of anomalies and create a sense of mistrust for the system. As a result, human attention and investigation are required. Similar to the risks encountered in other systems, malicious attacks on surveillance video content aim to hide content or change it to deceive viewers, such as security personnel and police applications.

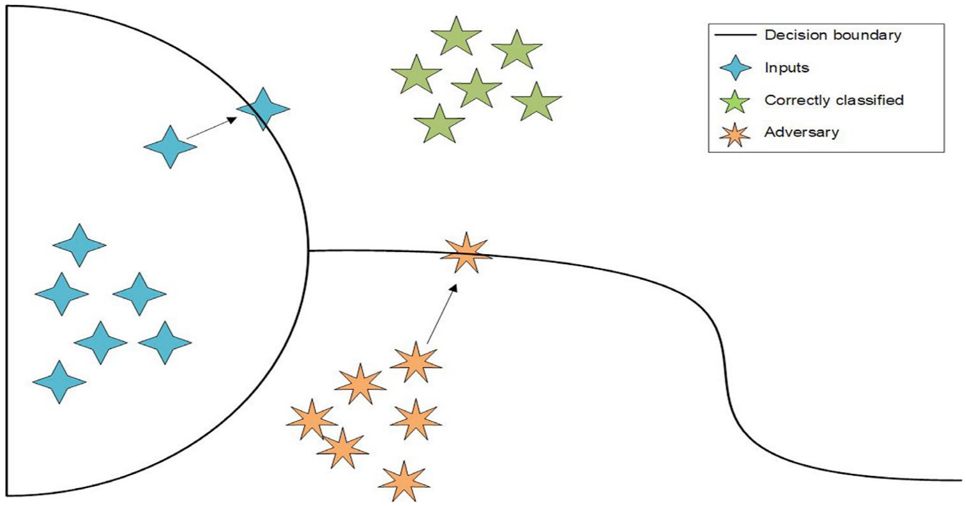


Figure 2.9 Illustration of an adversarial attack in the feature space

Image classifiers in ML have similar or higher accuracy levels than humans. Thus, using technology for surveillance is highly effective because surveillance is a repetitive approach that may produce errors if conducted by a human. If surveillance is conducted by technology, humans can focus on acting when something wrong happens (Ullah et al., 2020). Adversarial attacks can be divided into two classes, namely, black box and white box, on the basis of the extent of information accessible to the attacker (Li et al., 2019b). The level of risk associated with perturbations in real-world application is a concern that has not been addressed yet. The various methods of generating adversaries are given below, along with their advantages and disadvantages.

1. White-box attacks

The attacker has information about the system's architectural model, training details, weights and the instances in which the system was trained. The classifier function is vulnerable to attack because it is known to the attacker. In neural networks (such as FGSM and DeepFool), the backpropagation technique is used to attack because the gradients are known (Saxena 2020). In FGSM, the goal is to ensure that the model misclassifies the input image. The input image is given as *x* with label *y* in Equation 2.10, which generates the adversarial image (OpenAI, 2017).

$$Advx \ = \ x + \in . sign\left(\nabla xJ\ (\theta, x, y)\right), \tag{2.10}$$

where *x* is the input image, $Adv\ x$ is the generated adversarial example, *Y* is the actual label, ∈ is a multiplier to keep the perturbations small and unnoticeable, *θ* represents parameters and *J* is the loss. The gradients are decided on the basis of the input to ensure that the loss is minimal. Perturbation is added to each pixel value, contributing to the maximum loss, which is determined using the chain rule. An already trained system is fooled by such attacks, and misclassification occurs. A good example is given in Figure 2.11. The original image is distorted by adding a small perturbation, resulting in misclassification by the system labelling it as a gibbon image instead of a panda with a high level of confidence.

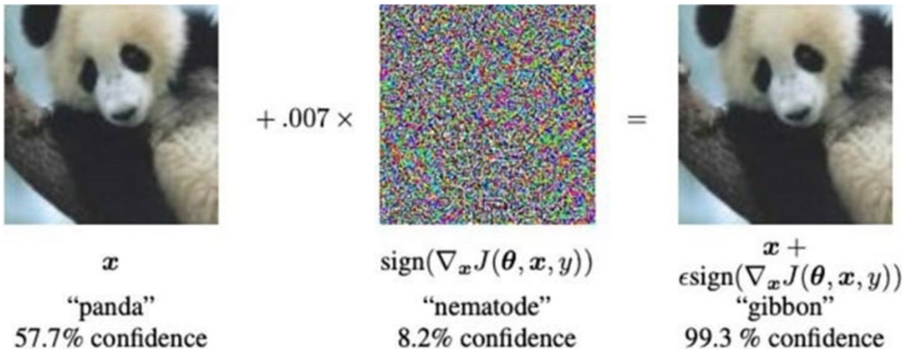

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure 2.10 Adversarial image generation (TensorFlow, 2020)

The advantages of generating an attack in a white-box setting are numerous. Generating an adversary image in this setting is easy because abundant information is available and accessible to the attacker. The attacker knows the model architecture, training data, the algorithm used and other crucial information. Another advantage is that less time is consumed, and the system's vulnerabilities become prominent. The disadvantages are that the white-box setting is inapplicable to real-life ML applications. Another disadvantage is that in the white-box setting, when an attack is constructed and the perturbations are made small to manipulate the input image, the computational cost increases.

2. Black-box attacks

Compared with white-box attacks, black-box attacks have access to the model's outputs but not its architectural model, weights, or other training details. Black-box attacks are conducted in several ways on the basis of the knowledge on outputs and other subareas, as follows.

A. Score-based attack

The output layer is accessible to the attacker, queries can be sent, and the final classification is available. With this technique, adversarial images in the black-box setting are developed using the confidence level, pixel value and loss. The confidence score is employed as feedback. The pixel values are increased, leading to high confidence. When the pixel value is reduced to decrease the confidence, misclassification occurs. A genetic-based algorithm is introduced to generate adversarial samples (Brendel et al., 2017).

B. Transfer-based attack

In this type of attack, a model is developed based on the knowledge and information available about the original system. An auxiliary or imitated model is created to replicate the operations and outcomes of the original system. This imitated model is then used for transfer-based attacks in a white-box setting. In this scenario, the attacker is aware of the training details, architecture and other parameters of the auxiliary model and utilises this information to target the imitated model. If the attack on the imitated model is successful, the same attack is then executed on the original model. Compared with score-based attacks, transfer-based attacks are more complex and time consuming because the attacker must first design and replicate the original system. Moreover, the computational cost is high for this type of attack because of the complexity of the operations that the auxiliary model must mimic.

C. Decision-based attack

The final output is accessible to the attacker. In this approach, the attacker utilises the random walk-on-the-boundary technique to search for adversaries (Dong et al., 2019).

Executing an adversarial attack in a black-box setting is beneficial because it applies to most real-life ML applications, such as self-driving cars. Another advantage is that in this setting, minimal knowledge is required, and the process is easier than that in transfer-based attacks. Another advantage is that this attack type is robust against defences compared with score-based and gradient-based attacks.

The disadvantages are that in a hard-label black-box setting, the generation of adversaries is unsuccessful. The random walk approach is used. However, this approach requires numerous queries and does not guarantee convergence, adding to the disadvantages. An optimisation-based technique has been developed to address the issues in the random walk approach, but the high dimensionality of the dataset and inputs makes the task difficult.

2.4.1 Generative Adversarial Network for Image Processing Applications

GAN is utilised in image processing applications mainly because of its success in generating images and videos (Peng et al., 2020b). In general, GANs are employed because they tend to improve the performance of previous methods (Du et al., 2020). In image processing, different areas, such as face detection, face recognition and facial feature extraction, have been explored using the DL method. GANs have recently garnered increasing attention because of their improved detection accuracy and reduced error rate. Face detection in image and video processing applications has been made easy by GANs. One of the benefits of using this network is that multiple GANs can be developed for different purposes, and they all work together to achieve the primary objective of the application (Liu et al., 2019). Aung et al. (2021) highlighted and proposed methods of improvement to the face detection process and system after combining the YOLO algorithm and VGG16 pre-trained CNN by utilising the DL Toolbox and the Image Processing Toolbox in MATLAB. Their experiment resulted in an increase in face detection speed in live, real-time video.

The human face identification method developed by Cuimei et al. (2017) utilises a Haar cascade classifier in conjunction with three other classifiers. By addressing variations in position, lighting and occlusions, the proposed method enhances the accuracy and reliability of face identification whilst reducing the number of false positives. The effectiveness of the suggested method was evaluated experimentally. The algorithm performs well in recognising human faces under various challenging settings, according to the results, and has potential applications in biometric systems, surveillance and human–computer interaction.

Many studies have developed a network of several GANs trained on a dataset to generate adversarial samples (Liu et al., 2018). According to Yu et al. (2018), GANs have flexibility in generating high-resolution images/videos, and they can work with other tools and techniques, providing easy integration and good prospects (Roheda et al., 2018). For example, Zhang et al. (2019a) developed a cross-GAN augmented with a variational autoencoder technique to address the problem of view disparity in redetecting and matching an already detected person and face. In addition, deep convolutional generative adversarial networks (DCGANs) (Aslan et al., 2019), IGAN-IDS (Huang & Lei, 2020) and the application of GAN for recognition (Bhat & Dharani, 2018) have been proposed.

Applying GAN to image processing applications has disadvantages and limitations. Given that traditional GANs are often difficult to train, a large image dataset can substantially improve the training time. GANs are also known to produce garbage values and nonsensical data. Deep convolutional GAN developed by Radford et al. [as discussed in (Donahue et al., 2018)] attempts to stabilise the training process of GAN by using several strategies.

The basic model of GAN is a complex concept. Therefore, several different variations have been developed. GANs are also popularly used for image-to-image translation, but they cannot preserve image details and information during translation according to Peng et al. (2020a). For facial expressions, according to Aggarwal et al. (2020) and Bozorgtabar et al. (2020), the practicality of GANs is limited.

Velastin et al. (2020) presented a computer vision methodology that utilises a standard video camera to accurately identify, monitor and enumerate individuals entering or exiting a metropolitan train. The method integrates background subtraction and foreground analysis approaches to identify individuals within video frames. The researchers utilised diverse computer vision techniques, such as image difference, connected component analysis and blob tracking, to accurately track and enumerate individuals. They performed experiments by using authentic video data obtained from a metropolitan train station to assess the efficacy of their methodology. The results of this study contribute to the area of video-based surveillance and offer valuable perspectives for enhancing crowd control and operational efficacy in urban railway stations.

Sun et al. (2019) introduced a new methodology for achieving resilient visual tracking by utilising a fusion of CNN and ELM. The proposed method employs CNN to extract features from the input data. The extracted features are subsequently inputted into ELM that functions as a classifier to estimate the object's position and track it across successive frames. The study presented a comprehensive exposition of the structure and execution of the CNN-ELM framework utilised in visual tracking. Furthermore, the study compared the proposed technique's outcomes with those of other advanced tracking algorithms. The results of this study contribute to the advancement of visual tracking methodologies and provide valuable insights for creating tracking systems that are dependable and effective in diverse contexts.

2.4.2 Generative Adversarial Network for Object Detection Applications

A survey of different object detection methods was conducted by Li et al. (2020a), who reported that CNN is an important model for DNNs. The problem of detecting moving objects by using CNNs was investigated by Zhu et al. (2020), who aimed to achieve real-time object detection. A novel framework that utilises coarse- and fine-grained techniques for detection was introduced. The moving regions in each frame were extracted using DCNNs, which made it possible for the regions to be precisely detected. High-resolution video frames were used to extract the features and train the system on video data. The results indicated that the proposed framework is more efficient and accurate than existing methodologies. The major drawback of the study is that the system was trained on high-resolution video frames, which is certainly not the case with video surveillance. The type of camera, lighting effects and other components affect the quality of video frames; such a framework is not favourable for use in a system that is installed for security purposes, according to Vähäkainu and Lehto (2019).

Table 2.3 provides a systematic review and comparison of the techniques discussed above. The table includes the techniques and algorithms used by different researchers for object detection and tracking in images and videos. The accuracy of each proposed model in object detection is demonstrated, along with its real-time object tracking performance on videos. Algorithm speed is also analysed by examining tracking speed using the frames per second (FPS) evaluation parameter. Furthermore, the performance of each model is evaluated based on the presence of various challenges, including occlusion, illumination, scale variance, multi-object detection and anomalies.

Table 2.3 Review of different object detection and tracking techniques in challenging scenes

	Technique	**Citation**	**Object Detection / Tracking accuracy**	**Total frames**	**Tracking speed**	**Performance in challenging scenes**
TLD	Improved TLD algorithm	(Zhen et al., 2020)	80%	500	26.32 fps	Performs well in the presence of a moving camera, motion blur, similar objects, scale change, illumination changes and occlusions
YOLO	with SORT algorithm	(Bathija & Sharma, 2019)	85.1%	930	–	
	YOLO-ACN	(Li et al., 2020b)	55.8% (average precision, AP)	4491	16 fps	Performs well in detecting occluded and small-sized objects in real-time videos
	with hard-example mining	(Ryu & Chung, 2021)	90.49%	–	–	Tested on images only. Detects objects accurately in the presence of occlusion
Adaboost	with Haar training	(Park et al., 2021)	85.9%	–	20 fps	Performs well in the presence of occlusions
CNN	Faster R-CNN	(Khan et al., 2019)	75.2%	–	25 fps	Performs efficiently in a real-time environment and detects multiple objects in a single frame
	with KCF	(Feng et al., 2018)	86.6% (Precision)	576	–	Detects human objects in the presence of occlusion and scale variance (SV)

	with correlation filters	(Yuan et al., 2020)	49% (59.7% SV, 59.2% occluded)	–	8 fps	The model is slow but effectively detects objects in the presence of occlusions and SV
GAN	NM-GAN	(Chen et al., 2021)	90.7%	–	0.031 fps	Detects objects in the presence of anomalies and detects anomalies as well
	OPGAN	(Ruiqiang et al., 2021)	84.2%	–	–	Tested on images only, efficient in detecting small objects
	DL with GAN	(Cheng et al., 2020)	91.9% (precision)	–	14.8 fps	Performs successfully in 11 challenging scenes, including occlusion, deformation and SV; the algorithm is slow, with poor tracking performance in the presence of long-term occlusion and similar objects
	GAN with Faster R-CNN	(Huang et al., 2018c)	89.65%	–	–	Detects small objects in images but has high complexity, is time consuming and requires extensive computation
	GAN-Do	(Prakash & Karam, 2021)	67.47%	–	–	Addresses the problem of object detection in reduced-quality images; performs well in detecting images in settings with camera-shake blur, Gaussian blur, defocus blur and additive white Gaussian noise

According to Zhu et al. (2020), DCNNs are used for object identification. Traditional approaches often struggle to reliably detect moving objects in complex scenarios because of difficulties, such as varying lighting conditions, object motion and occlusions. The authors suggested a method to boost the effectiveness of moving object identification by utilising DCNNs. They built an innovative DL architecture especially for this task; this architecture can learn discriminative features and capture the spatiotemporal data contained in video sequences. The effectiveness of the suggested strategy was evaluated experimentally to demonstrate its advantages in terms of accuracy and resilience. The results have ramifications for several applications that depend on accurate and immediate object recognition.

Ye et al. (2018) proposed a DL-based methodology to identify and monitor mobile entities by utilising a solitary camera installed on UAVs. The proposed methodology employs CNNs, a type of DL model, to acquire discriminative features from the frames of aerial videos. The study presented empirical findings that proved the efficacy of the proposed approach. Various measures, such as accuracy, resilience and computational efficiency, were adopted. The authors discussed the algorithms considered for practical applications in UAV systems, such as surveillance, search and rescue operations and environmental monitoring. The study's results contribute to the domain of computer vision using UAVs. The findings offer important perspectives for the advancement of reliable and sophisticated UAV systems that can be utilised in diverse contexts.

The advantages of using GANs for object detection include the fact that it does not require paired input information, and using a simple mask mean loss function can enhance GANs' performance in detecting objects. GANs are a powerful learning model that can generate real-like images in high resolution without requiring supervised learning, as stated by Lee et al. (2018). However, supervised learning can be achieved for other purposes, such as object discrimination in space, according to Wu et al. (2019).

GANs require a normal sample and can detect even the smallest defect (Zhong et al., 2020). Face recognition in surveillance videos has been used in addition to object detection (Bhat & Dharani, 2018). GANs are preferred for their ability to be trained on adversarial samples and detect anomalies. Once a GAN is successfully trained, it can generate a large amount of data with minimal computational requirements. Peng et al. (2020a) introduced a technique for enhancing the efficacy of facial recognition across diverse domains through the use of a soft semantic representation approach. The method employs advanced DL techniques to acquire distinctive features and semantic representations from facial images. Moreover, it introduces a soft assignment approach to promote the resilience of the acquired representations to domain shifts and variations. The efficacy of the proposed method was demonstrated through experimental assessments on benchmark datasets, and it surpassed current methods in terms of accuracy in cross-domain face recognition. The study also examined the interpretability and robustness of the acquired soft semantic representations, along with potential real-world applications of the suggested approach. The findings suggested that utilising soft semantic representations may have the capacity to effectively resolve the difficulties associated with cross-domain face recognition, thereby propelling the domain of biometric security forward.

Although GANs have many advantages, they still have limitations. Their algorithms have two parts: a generator and a discriminator. The two networks are difficult to train concurrently, posing a threat to their practical use in detecting object position-based anomalies (Yang et al., 2019b; Alnujaim et al., 2019). Using GANs for hand gestures is crucial to computing similarity because their absence is likely to increase failure during the training process. Wang et al. (2020c) developed an evaluation metric to judge the performance of GANs. They revealed that GANs are challenging, requiring large datasets for training, and they cannot generate images of human-perceived quality.

A method for real-time recognition and tracking of tiny target traffic signs in a video footage was presented by Song et al. (2021). For accurate and efficient results, the proposed method combines object identification and tracking techniques. CNN and other DL techniques are used to provide accurate object recognition, and tracking methods are adopted to retain target identity over several video frames. The suggested framework, along with the architecture of the object detection network and the tracking method, was thoroughly described. The authors discussed the practical ramifications of their real-time detection and tracking technique. They emphasised its potential use in intelligent transportation systems, where timely and accurate recognition of small target traffic signs is essential for maintaining traffic flow and road safety. The results have substantial implications for intelligent transportation systems that aims to enhance traffic control and road safety.

An innovative method for quick and reliable object tracking in video sequences was presented by Shin et al. (2020). The suggested approach addresses the difficulties in maintaining accurate object tracking in complex and dynamic contexts by combining KCF with a tracking failure warning technique. The tracking failure detection module checks tracking performance and initiates re-detection or recovery techniques as needed. The KCF algorithm is used for its computing efficiency and resilience in managing appearance changes. The authors conducted experimental assessments to evaluate the efficiency of their suggested method for object tracking. They found that their technique offers a quick and reliable solution to object tracking and has several useful applications in computer vision. This advancement in computer vision technology provides an efficient solution for object tracking.

2.4.3 Generative Adversarial Network for Behaviour Detection

Video surveillance systems are developed for different purposes, such as behaviour detection, object detection, crowd estimation, human identification and face recognition (Li et al., 2017). Video surveillance mostly produces crowded images. For this reason, many scholars have used CNN with GAN and other techniques or functions to improve performance. Examples include CS-GAN for estimating crowd density (Zhang et al., 2020a), FCGAN to produce quality images of faces (Huang et al., 2018c) and contour GAN (Yang et al., 2019a).

GANs can be employed to solve the resolution problem, especially when the footage contains people and obstructions found in surveillance. Conditional GANs are also beneficial for solving the problem of missing modalities (Roheda et al., 2018). However, the application of GANs in behaviour detection and recognition has limitations. For instance, GANs have poor generalisation ability possibly because their models cannot fully learn the distribution of the dataset. GANs are also unable to detect untrained gestures, leading to their limited in application in novelty detection. The GAN framework was evaluated by Simao et al. (2019), who revealed that quality images are generated by the discriminative model, but the baseline neural network fails to detect untrained gestures.

The advancement in the global world, the use of social media platforms and investment in digital currencies influence DL approaches, and many researchers have used CNN along with other parameters to understand the socio-economic effects on price prediction, especially in terms of bitcoin. Aggarwal et al. (2019) utilised root mean square error (RMSE), LSTM, CNN and GRU to conduct a comparative study between bitcoin and gold prices. They concluded that DL algorithms are efficient in predicting bitcoin prices, with LSTM yielding the lowest RMSE value. However, analysis of sentiments on Twitter showed a positive correlation between optimistic and pessimistic tweets, with the extent of the influence depending on the popularity of the user, indicating a remarkable association between the occurrence of optimistic tweets concerning bitcoin and an anticipated increase in its prices.

2.5 Uses of GAN in Video and Image Accuracy and Detection

GAN is an AI model comprising two key components: a generator network and a discriminator network. GAN is widely utilised in visual media, including videos, images and graphics, because of its ability to produce authentic and superior-quality output. Goodfellow et al. (2014a) introduced a generative modelling method that utilises the framework known as GAN. It comprises two fundamental components, namely, a generator network and a discriminator network. The two networks are trained concurrently in an adversarial manner. The training process involves a minimax game, wherein the generator network endeavours to minimise the discriminator's capability to distinguish between authentic and synthetic samples, and the discriminator network strives to maximise its discriminatory potential.

The authors explicated the theoretical foundations that encompass the utilisation of Jensen–Shannon divergence as a metric for assessing the resemblance between the distributions of the generated and authentic data. The study delved into diverse implementations of GANs, including image synthesis, style transfer and data augmentation. It underscored the capability of GANs to comprehend intricate patterns and produce visually captivating outputs of superior quality. Hence, GANs are robust frameworks for generative modelling, offering valuable insights.

Radford et al. (2015) examined the utilisation of DCGANs in unsupervised representation learning. The architecture under consideration comprises a generator network and a discriminator network, both of which are founded on CNNs. The authors presented a set of architectural principles aimed at ensuring the stable training of DCGANs. These principles included the exclusion of fully connected layers in favour of convolutional layers, the incorporation of batch normalisation, the utilisation of specific activation functions and the implementation of stridden convolutions.

The authors conducted experiments on various datasets, including ImageNet, to assess the efficacy of DCGANs, which exhibited the ability to produce lifelike images with a range of distinctive variations, implying the efficacy of the acquired representations. Furthermore, the interpretability of DCGANs through the execution of arithmetic operations on the acquired representations has enhanced the comprehension of profound generative models and their capability to acquire substantial representations from unannotated data, thereby paving the way for further progress in unsupervised learning methodologies.

Karras et al. (2017) proposed a novel approach called progressive growing, which involves training GANs to produce images of superior quality and diversity. A method of addressing challenges, such as mode collapse, instability, and restricted diversity in generated samples, was suggested by gradually increasing the size of the generator and discriminator networks during the training process. The study provided empirical evidence that supports the efficacy of progressive growing and achieved noteworthy enhancements in image quality, stability and diversity compared with conventional GAN training techniques. The authors discussed supplementary methodologies, including minibatch standard deviation and equalised learning rate, to augment the efficacy and consistency of the progressive growing methodology. They elucidated the potential usages of progressive growth in producing high-resolution images, investigating image synthesis and enhancing data augmentation methodologies. The constraints and difficulties associated with the approach, such as computational demands and the necessity for meticulous calibration, were also examined.

Considering the importance of GAN in video and image accuracy and precision, Wang et al. (2018) introduced a framework that utilises a combination of GAN and attention mechanisms to produce videos based on textual descriptions. The framework comprises two distinct components, namely, a text-to-video generator and a discriminator. The process of generating videos from text involves the utilisation of an RNN that employs an attention mechanism to decode textual descriptions and convert them into corresponding video frames. The network responsible for discrimination is trained to differentiate between authentic videos and artificially generated ones. The efficacy of the authors' methodology was exhibited through experimentation on established benchmark datasets, including MSR-VTT. The videos exhibited a notable degree of visual fidelity and consistency with the provided textual depictions. The study also examined plausible uses of video creation from textual input, including the generation of video captions, production of virtual reality content and provision of support for video editing. Despite its utility, the framework has certain limitations, including its dependence on pre-segmented videos and difficulty in producing videos that are lengthy and intricate.

In continuation to the abovementioned concept, Zhu et al. (2017) introduced a new methodology for accomplishing unpaired image-to-image translation through the use of CycleGAN. The framework comprises a pair of mapping functions and two adversarial discriminators that engage in an adversarial game during the training phase to acquire significant mappings between the two domains. The study showcased empirical findings on diverse image translation endeavours, including style transfer, object transfiguration and season conversion. The images produced exhibited effective domain adaptation and accurate translation even in the absence of paired training data. The researchers investigated the transferability of the acquired models, demonstrating the capability to utilise a pre-existing CycleGAN that has undergone training for various image translation assignments. The study expounded on the merits and demerits of CycleGAN, which included obstacles pertaining to mode collapse and the arduousness of regulating the precise features of the converted images. The framework has potential applications in various domains, including artistic creation, image editing and data augmentation.

In general, CycleGAN is a robust framework for unpaired image-to-image translation, facilitating progress in the domain of unpaired image synthesis and manipulation. In line with the importance of GAN for images, Vid2Vid was introduced by Wang et al. (2018) as a framework for the visualisation of human activities through video-to-video translation. The proposed approach utilises a fusion of conditional GANs and an optical flow-based warping module to attain superior video translation outcomes. The authors proposed a biphasic training approach for Vid2Vid, where a network for pose-to-pose translation is trained to produce intermediary human poses from the source video. To improve the authenticity and consistency of the synthesised videos, the researchers utilised a warping module based on optical flow. This module aligned the generated frames with the intended pose and motion of the target. The efficacy of Vid2Vid was exhibited through experimentation on diverse human activity datasets, where Vid2Vid showed superior visual fidelity, seamless motion and authentic human activity depiction.

Thorough examination of the aforementioned studies indicates that GAN can produce images that exhibit a high degree of realism and are reminiscent of a specific style or category. By training the generator network on a vast dataset, GAN can generate novel images that exhibit comparable characteristics and structures. Furthermore, it is highly beneficial in the fields of computer graphics, art and design. It can create innovative and aesthetically pleasing images.

GAN can be utilised for various image editing tasks, including colorisation, inpainting and super-resolution. It enhances image quality by identifying and learning the underlying patterns and structures present in the training data. As a result, GAN can effectively fill in the missing parts of an image or improve its resolution. The process of style transfer involves the utilisation of GAN to acquire the stylistic features of a given image and subsequently applying them to another image. The use of this methodology facilitates the generation of visually captivating and distinctive visuals through the amalgamation of the substance of one image with the style of another.

GAN has been employed for style transfer in various domains, including artistic filters and photo manipulation. It is effective in generating video sequences that are realistic and coherent. By extending image generation principles to the temporal domain, GAN can produce videos frame by frame. It has potential applications in various fields, such as video editing, special effects and virtual reality. It can be utilised to produce synthetic videos or modify pre-existing footage.

The GAN training process enables the prediction of forthcoming frames within a given video sequence. By acquiring knowledge on the dynamics and patterns inherent in the training data, GAN can produce credible future frames.

2.6 Loss Function

In ML and optimisation, a loss function, which is also referred to as a cost or objective function, is a mathematical construct that quantifies the difference between the anticipated output of a model and the actual output. The process of quantifying the error or loss that is associated with the predictions made by a model is a crucial aspect of ML. The primary objective is to minimise this loss during the training phase.

Types of Loss Function: Various loss functions are available, and the selection of an appropriate one is contingent upon the particular problem and the characteristics of the data. The following text shows frequently utilised loss functions.

- Mean squared error (MSE) is a loss function that computes the mean of the squared differences between the predicted and actual values. The approach is frequently employed in regression tasks and imposes a substantial penalty on large errors.
- Mean absolute error (MAE) is a metric used to quantify the average absolute deviation between the predicted and actual values. The method exhibits a higher degree of robustness in the presence of outliers compared with MSE and is frequently employed in regression applications.
- The binary cross-entropy metric is commonly employed in binary classification tasks, where the resulting output is restricted to binary values of 0 or 1. The aforementioned statement pertains to the quantification of the degree of difference between the predicted probabilities and the actual binary labels.
- The categorical cross-entropy method is utilised in scenarios involving multi-class classification tasks. This metric quantifies the disparity between the anticipated probabilities of class membership and the actual class labels.

- Kullback–Leibler divergence is a mathematical measure that quantifies the dissimilarity between two probability distributions. It is frequently employed in various applications, including generative modelling and reinforcement learning.
- The hinge loss function is frequently used in SVMs and is well-suited for binary classification tasks. The objective is to optimise the margin between different classes.
- The *Huber* loss function is a hybrid of MSE and MAE functions. It has a higher degree of resistance to outliers compared with MSE and achieves harmonious equilibrium between robustness and smoothness.

These examples of loss functions are merely a subset of the broad range of available options, with the possibility of developing tailored loss functions to suit particular objectives and specifications. The selection of an appropriate loss function is contingent upon the specific problem being addressed and the intended attributes of the model's prognostications.

The importance of loss functions to toxicity probability interval (TPI) design was expounded by Siegel (2021). Loss functions are utilised to measure the difference between the estimated probabilities and the target toxicity rates. These functions substantially affect the decisions made during the process of dose escalation or de-escalation. The study employed simulated and empirical illustrations to showcase the influence of distinct loss functions on the efficacy of TPI design. Frequently employed loss functions include quadratic, absolute difference and logarithmic loss functions. The utilisation of adaptive loss functions provides a degree of adaptability and facilitates the integration of progressive insights into the dose-finding process. The study highlighted the importance of meticulous selection and assessment of loss functions and other statistical methodologies in TPI design to enhance patient safety and treatment outcomes.

Hoffer and Ailon (2015) investigated the utilisation of triplet networks in the context of deep metric learning. The authors suggested a theoretical construct that adopts triplet loss to acquire proficient representations within a DL context. The triplet loss function is predicated on the use of triads of training samples, which comprise an anchor, a positive instance from the same category as the anchor and a negative instance from a distinct category. The authors conducted experiments on benchmark datasets to showcase the efficacy of triplet networks in the context of metric learning tasks. The results demonstrated that the acquired representations manifest superior discriminatory capacity and have the potential to augment efficacy across diverse recognition assignments. The study also examined various facets of triplet network training, including the identification of informative triplets, batch construction methodologies and network architecture selections. It provided methods of mitigating the computational difficulties that arise when training triplet networks on a large scale. Ultimately, the authors conducted a comparative analysis between their proposed methodology and alternative metric learning techniques, namely, Siamese networks and contrastive loss. They emphasised the superior performance of triplet networks and their capability to manage extensive datasets.

The utilisation of Dice loss was suggested by Li et al. (2019a) as a potential solution for addressing data imbalance in natural language processing (NLP) assignments. The Dice loss function is a mathematical construct that is based on the Dice coefficient, which is a widely employed metric for assessing the degree of resemblance between sets. The authors conducted experiments on various NLP datasets to showcase the efficacy of Dice loss in enhancing the performance of imbalanced tasks. The authors compared Dice loss with other commonly used loss functions, including cross-entropy and focal losses. Their findings indicated that Dice loss exhibits superior performance in managing class imbalance and generating precise predictions for underrepresented classes. The study analysed the theoretical underpinnings of Dice loss and elucidated its benefits, including its superior management of class imbalance, enhanced gradient propagation and ability to concentrate on particular classes of interest. To summarise, the study presented Dice loss as a potentially effective remedy for NLP tasks that suffer from imbalanced data. The results provide substantial contributions to the field of imbalanced dataset research in NLP and can be beneficial for scholars and professionals.

The challenges related to training deep feedforward neural networks were investigated by Glorot and Bengio (2010). The objective was to elucidate the variables that make deep networks more challenging to train in comparison with shallow networks. The authors addressed the vanishing gradient and exploding gradient phenomena, which impede the efficient training of DNNs. A novel initialisation method called normalised initialisation was introduced to maintain gradient magnitude stability across the network. The researchers engaged in a discourse regarding the importance of activation functions in the training of deep networks. They considered the implications of various network architectures, including convolutional and recurrent networks, and the challenges encountered during training. They offered notable perspectives towards the challenges associated with the training of deep feedforward neural networks and presented pragmatic approaches to enhance the training procedure. The results enhance profound neural networks, propelling the domain of profound learning.

Zancato et al. (2022) investigated the notions of trainability and generalisation in DNNs. They conducted a comprehensive study on the trainability and generalisation of deep networks by utilising theoretical analysis and empirical experiments to explore the various factors that influence these aspects. They engaged in a discourse regarding the obstacles that arise when training deep networks, including the emergence of vanishing or exploding gradients, overfitting and optimisation complexities. They reviewed various strategies that can be employed to resolve these challenges. Furthermore, the study explored the notion of generalisation and scrutinised the variables that affect a network's capability to generalise effectively to unobserved data. The results have substantial implications for optimising the training process, augmenting generalisation abilities and propelling the domain of DL forward (Li et al., 2023).

2.7 Synthesis of Related Work

In this section, a summary of the most relevant research from the literature presented in the previous sections and other focused studies is provided. Some of the references mentioned earlier are cited again for further analysis and discussion of the findings. Analysis of related studies, such as those of Kalbo et al. (2020), Senthil Murugan et al. (2018), Qiu et al. (2019) and Zhang et al. (2020a), revealed that video surveillance systems are vulnerable to adversarial attacks. ML techniques, particularly those developed using DL to automate tasks (e.g. face detection, object detection and weapon detection), have been found to be particularly susceptible to these attacks. Extensive research has been conducted on the subject, demonstrating the severity of the problem. One of the reasons is that evaluation of the efficacy of such techniques is often difficult because ML algorithms mostly learn from factors that are not necessarily malicious, leaving a gap for accurate detection. Another emerging threat is using or spying on surveillance systems in some external networks. Analysis of some of these research revealed missing contributions for networks that include automatic detection of malicious behaviour.

Some studies, such as that of Goodfellow et al. (2014b), developed GAN models to work with images, and others focused on specific detection problems. However, no study has evaluated object detection and human malicious behaviour in a real-time environment. A detailed comparison of past studies is presented in Tables 2.4–2.7 and Tables 2.8–2.10.

Table 2.4 Comparison of literature review findings

Study	Focus	Research Findings	Gaps
(Goodfellow et al., 2014c)	Addressed the problem of adversaries by using AI approaches	A two-player generator - discriminator model to generate and identify fake data from real data	Used images only; it has room for improvement in terms of accuracy and efficiency
(M. Divya, 2016)	Used the Global GIST feature to detect a moving human body in the video	Detection	Limited to detecting human movement and does not classify or recognise malicious behaviour; works on video data but overlooks the problem of tampering
(Zhang et al., 2019b)	Detecting tampered image	Obtains an F1 score of 0.98, good performance	Detects tampered images but cannot be translated to video processing

(Zhang et al., 2020b)	Scaling low-resolution images to super resolution	Achieves easier detection and better performance than others in detecting small objects	Does not address the problem of object detection in real time or the issues of tampering

Table 2.5 Summary of literature review findings by Nguyen et al. (2023a)

(Nguyen et al., 2023a)	
Key Concept	Surveillance system, privacy policy, targeted advertising, DL
Article Type	Review paper
Methodology	**Contribution:** Proposed a framework to analyse physical adversarial attacks and provided a comprehensive survey of physical adversarial attacks on four key surveillance tasks, namely, detection, identification, tracking and action recognition, under this framework. **Method:** The authors described several methods employed in physical adversarial attacks, including the use of adversarial patches, stickers, glasses, hats, masks and makeup, to deceive surveillance systems. These methods involve optimising adversarial patterns or perturbations to minimise the detection or recognition accuracy of the targeted systems. The optimisation considered task loss (e.g. face recognition accuracy) and physical loss (e.g. smoothness or realism of the adversarial pattern). The study also categorised different attack methods on the basis of their objectives and provided detailed explanations of each method.

Findings	**Adversarial Patches:** Adversarial patches were successful in hiding persons from the YOLO detector. The detection accuracy was reduced from 100% to 17% by using eight patches. The same patches also decreased the detection accuracy of the SSD detector from 75% to 13%. **Adversarial T-Shirts:** Adversarial t-shirts were able to hide persons from the Faster R-CNN detector, reducing the detection accuracy from 100% to 17% by using eight patches. These t-shirts also decreased the detection accuracy of the YOLO detector from 100% to 69% and that of the SSD detector from 75% to 13%. **Adversarial Glasses:** Adversarial glasses were able to fool the state-of-the-art facial recognition system ArcFace, making the wearer essentially invisible to the system and even tricking it into thinking they are someone else. **Physical Adversarial Textures (PATs):** The adversarial learning process for PATs involved iteratively updating PAT by backpropagating the tracking loss from the tracker. The physical adversarial poster method successfully misled the human tracker GOTURN with a fooling ratio of 93%. Over-the-Air Adversarial Flickering Attacks: This method misled human action recognition models with a fooling ratio of 93%. **Metrics:** The metrics used included fooling ratio, mean absolute perturbation per pixel (MAPer), and mean absolute temporal-difference perturbation per pixel.
Interpretation	The results of various experiments showed that physical adversarial attacks can successfully deceive surveillance systems, leading to misclassification or misidentification of individuals. However, challenges persist in creating robust and transferable physical adversarial attacks. These challenges include the need to consider real-world conditions, such as lighting, viewing angles and camera distortions, as well as the limitations of current optimisation algorithms. Future work should focus on addressing these challenges, exploring attacks on video-based tasks, investigating attacks on different imaging spectra (such as infrared) and developing highly effective defence mechanisms against physical adversarial attacks in surveillance.
(Wei et al., 2018)	
Key Concept	Generative adversarial network
Article Type	Experimental application

Methodology	**Contribution:** The study presented a novel method, unified and efficient adversary (UEA), for attacking image and video object detection models. It is the first of its kind to efficiently manage both images and videos whilst simultaneously fooling proposal- and regression-based detectors. UEA also introduces a multi-scale attention feature loss to enhance its black-box attacking ability by manipulating the feature maps from multiple layers to deceive detectors. UEA offers better transferability compared with existing attacking methods, thereby extending its ability to attack a broad range of object detectors. **Method:** UEA formulates the problem within the conditional generative adversarial network (GAN) framework. Here, a generator network is trained to produce adversarial images and videos. The method incorporates a high-level class loss and a low-level feature loss into the GAN framework to jointly train the generator. The generator network is structured as an encoder–decoder network with 19 components, and the discriminator network is similar to ResNet-32. The authors also utilised adversarial networks to generate adversarial examples for semantic segmentation and object detection and employed flow-guided feature aggregation and deep feature flow for video recognition.
Findings	UEA achieved a substantial accuracy drop for Faster R-CNN and SSD300 object detection models. Specifically, it resulted in a 0.65 accuracy drop for Faster R-CNN and a 0.48 accuracy drop for SSD300. This performance is 12 times better than that of the previous method known as DAG. In terms of efficiency, UEA was approximately 1,000 times faster than DAG in generating adversarial examples, demonstrating its superior speed. In video object detection, UEA achieved a 0.40 mean average precision (mAP) drop for Faster R-CNN and a 0.44 mAP drop for SSD-300. These findings indicate that UEA is effective in reducing the precision of these models in detecting objects within videos.
Interpretation	The experimental results revealed that UEA outperforms the existing attacking method DAG in terms of attacking Faster R-CNN and SSD300 detectors. UEA achieves a large mAP drop and generates imperceptible adversarial examples, making it much faster than DAG in producing adversarial examples. The future work direction or challenge mentioned in the study involves exploring the attacking performance of UEA on advanced object detectors, such as those using ResNet 101 as the backbone network and FCN as the object detector. Further research could also be conducted to improve the mAP drop and explore the transferability of UEA to other types of object detection models. Future research should focus on generating robust adversarial examples and understanding the properties of neural networks.
(Zheng et al., 2020a)	
Key Concept	Cameras, dispersion, image colour analysis, iterative methods, neural networks
Article Type	Experimental application

Methodology	**Contribution:** The study introduced an effective adversarial attack on person re-identification in video surveillance systems known as the dispersion reduction (DR) attack model. This novel approach generates adversarial images that substantially degrade the performance of state-of-the-art person identification models. Reducing the dispersion of the internal feature map of a neural network makes the objects within the feature map less recognisable and distinguishable, highlighting potential security risks in video surveillance systems. **Method:** The DR attack model works by iteratively reducing the dispersion of the internal feature map of a neural network to generate adversarial examples. The method updates these adversarial examples on the basis of the dispersion gradient and keeps the perturbations within a set perturbation budget. Three different state-of-the-art person-identification models were used as victim models and attacked using this DR attack model.
Findings	**Baseline Performance:** The mean average precision (mAP) scores for the baseline victim models (DG-Net, AlignedReID and PLR-OSNet) on various datasets were as follows: DG-Net: 86.0% (Market-1501), 61.1% (CUHK03), 74.8% (DukeMTMC-ReID), 52.3% (MSMT17) AlignedReID: 82.3% (Market-1501), 70.7% (CUHK03), 82.8% (DukeMTMC-ReID), 43.7% (MSMT17) PLR-OSNet: 88.9% (Market-1501), 77.2% (CUHK03), 81.2% (DukeMTMC- ReID) **Effect of Attacks:** After implementing different attack methods, the drop in mAP values was as follows: TI-FGSM: caused a drop in mAP values ranging from 23.6% to 58.1% TI-DIM: caused a drop in mAP values ranging from 29.4% to 67.7% Proposed DR attack: caused the most notable drop in mAP values ranging from 7.8% to 9.5%
Interpretation	The experimental results demonstrate the effectiveness of DR attack. The adversarial examples generated by this method substantially affect the person’s re-identification performance even though the perturbations are imperceptible to the human eye. DR attack outperforms other state-of-the-art attack models and leads to the most notable drop in mAP values. **Challenge:** Despite its effectiveness, the DR attack model still faces challenges concerning robustness against defence mechanisms and scalability to large-scale datasets. According to the study, future work should explore defence mechanisms against adversarial attacks in person reidentification and develop advanced attack strategies to overcome these challenges.
(Li et al., 2021b)	
Key Concept	Adversarial attack, deep neural network (DNN), object detection
Article type	Original development

Methodology	**Contribution:** The study presented a novel bidirectional adversarial attack approach against deep neural network (DNN)-based object detectors to expose the security vulnerabilities of these models and raise awareness about the potential misuse of object detection techniques. This method includes the generation of adversarial examples that can mislead the object detection model by minimising the background probability and maximising the ground-truth probability. It also introduces a new confidence loss function to decrease foreground scores in the region proposal network. **Method:** The proposed bidirectional adversarial attack approach consists of two steps: generating adversarial examples and launching adversarial example attacks. The adversarial examples are created using a pretrained autoencoder as the generator, and the model is trained with an adversarial approach to improve the similarity between the adversarial examples and the original images and to hasten algorithm convergence. These adversarial examples are then used to attack the object detection model by optimising the perturbations. This attack misleads the model towards the background class and away from the ground-truth class. A confidence loss function is also applied to reduce the foreground scores in the region proposal network.
Findings	The proposed bidirectional adversarial attack approach achieved a substantial 58.3% mean average precision (mAP) drop on the Faster R-CNN object detection model in the white-box attack scenario. In the black-box attack scenario, the adversarial examples generated by the proposed method resulted in a 39.5% mAP drop when applied to the YOLOv3 object detection model. These results demonstrate that the proposed bidirectional adversarial attack approach effectively attacks the region confidence and classification accuracy of various object detection models. They prove the strong concealment and satisfactory transferability of the adversarial examples generated by this method, which can achieve substantial interference even in unknown object detection networks with small perturbations that are inconspicuous to humans.
Interpretation	The proposed method achieves a notable drop in mAP on the object detection models. The adversarial examples generated by this method demonstrate superior transferability because they can also attack different object detection frameworks. **Challenges:** The challenge lies in making the generated patches more inconspicuous and less noticeable to humans. Future work could focus on addressing this issue and further improving the robustness and stealthiness of adversarial attacks against object detectors. Moreover, the attack's effectiveness can still be improved, and possible defences against adversarial attacks on object detection models should be explored.
(Thys et al., 2019b)	
Key Concept	Detectors, image colour analysis, cameras, robustness, optimisation, computer vision, surveillance

Article Type	Optimisation
Methodology	**Contribution:** The study contributed an innovative approach for generating adversarial patches specifically designed to fool person detectors, thereby highlighting potential vulnerabilities in automated surveillance systems. The objective was to create a patch that can effectively conceal a person from a person detector, which could potentially be exploited to bypass security systems. **Method:** The authors employed an optimisation process to generate these adversarial patches. They focused on minimising three parameters: the non-printability score, the total variation in the image and the objectless score. These patches were trained using actual images of different individuals and applied to these images under various transformations to enhance their robustness.
Findings	PR curve analysis demonstrated that the OBJ approach substantially lowers the accuracy of person detection. However, the exact numerical value is not specified in the provided context. In terms of recall percentages, the OBJ-CLS approach has the lowest recall percentage amongst the compared methods, indicating that it is the most effective in circumventing alarms generated by surveillance systems. Again, the exact numerical value is not specified. Figures 6 and 7 in the paper provide visual evidence of the patch's effectiveness in hiding persons from the object detector in digital and printed versions. The specific numbers or percentages related to this effectiveness are not provided in the given context.
Interpretation	The results demonstrate that the patches generated substantially reduce the accuracy of person detection, as evidenced by the PR curve analysis. In most cases, the patches successfully conceal individuals from the detector. **Challenges:** The primary challenge encountered in this study is the wide variability in human appearances, different contexts and the absence of a consistent location to place the patch. The authors stated that future work could concentrate on enhancing the robustness of the patches through additional transformations and improving their transferability across different architectures.
(Thakur & Li, 2022b)	
Key Concept	Training, DL, perturbation methods, surveillance, Gaussian noise, conferences, neural networks
Article Type:	Experimental application

Methodology	**Contribution**: The study introduced a novel method called pseudo-adversarial training (PAT) designed specifically for detecting adversarial videos that have been manipulated with the intention of deceiving video analysis algorithms. The primary contribution lies in the development of this new approach, which is geared towards enhancing the robustness of video analysis algorithms against adversarial attacks. **Method:** The PAT method incorporates a training technique that utilises real and adversarial videos. The process generates adversarial videos through a pseudo-adversarial attack, a modified form of the traditional adversarial attack. These adversarial videos are then used to train the video analysis model, bolstering its resistance to adversarial attacks. The approach also includes a regularisation term to further reinforce the model's robustness.
Findings	Experiments on UCF-101 and Jester datasets demonstrated that the approach is highly accurate in detecting the adversarial input produced by different attacks. PAT achieves an AUC of 94.2% on clean and adversarial (containing both sparse and dense attacks) videos.
Interpretation	Experimental results demonstrated that the PAT method outperforms existing methods in detecting adversarial videos. It effectively identifies various types of adversarial videos, including those manipulated to deceive specific video analysis tasks, such as object detection and action recognition. **Challenge:** One of the key challenges identified is the dynamic nature of adversarial attacks. Attackers can continuously modify their techniques to avoid detection, making it challenging to keep pace with their evolving strategies. Future research could focus on developing advanced detection methods that can handle these evolving adversarial attacks. The study also suggested the potential application of PAT to areas beyond video analysis, including image classification and natural language processing.

This part addresses other related work that motivated the present research. Nguyen et al. (2023b) proposed a framework for the analysis of physical adversarial attacks and provided a comprehensive survey of such attacks on four key surveillance tasks: detection, identification, tracking and action recognition. They described several methods employed in physical adversarial attacks, including the use of adversarial patches, stickers, glasses, hats, masks and makeup, to deceive surveillance systems. Various experiments demonstrated that physical adversarial attacks can successfully deceive surveillance systems, resulting in the misclassification or misidentification of individuals. However, challenges persist in creating robust and transferable physical adversarial attacks. These challenges include the need to consider real-world conditions, such as lighting, viewing angles and camera distortions, as well as the limitations of current optimisation algorithms. For instance, one of the main findings of this study is that adversarial patches are successful in hiding persons from the YOLO detector, reducing its detection accuracy from 100% to 17% by using eight patches. The same patches also reduce the detection accuracy of SSD from 75% to 13%. The metrics used in many of the reported studies included fooling ratio, mean absolute perturbation per pixel and mean absolute temporal-difference perturbation per pixel. One of the future research directions mentioned in this work is developing effective defence mechanisms against physical adversarial attacks on surveillance systems.

Wei et al. (2019) presented a novel method called unified and efficient adversary (UEA) for attacking image and video object detection models. The UEA method offers better transferability compared with existing attacking methods, extending its ability to attack a broad range of object detectors. UEA formulates the problem within the conditional GAN (cGAN) framework. Here, a generator network is trained to produce adversarial images and videos. The method incorporates a high-level class loss and a low-level feature loss into the GAN framework to jointly train the generator. The generator network is structured as an encoder–decoder network with 19 components, and the discriminator network is similar to ResNet-32. In addition, the authors used adversarial networks to generate adversarial examples for semantic segmentation and object detection and employed flow-guided feature aggregation and deep feature flow for video recognition. The UEA method achieved a notable accuracy drop for Faster R-CNN and SSD300 object detection models. Specifically, it resulted in a 65% accuracy drop for Faster R-CNN and a 48% accuracy drop for SSD300. This performance is 12 times better than that of the DAG method. In terms of efficiency, UEA is approximately 1,000 times faster than DAG in generating adversarial examples, demonstrating its superior speed. In video object detection, UEA achieved a 0.40 mean average precision (mAP) drop for Faster R-CNN and a 0.44 mAP drop for SSD300, indicating that UEA is effective in reducing the precision of these models in detecting objects within videos. For future work, the authors recommended exploring the attacking performance of UEA on advanced object detectors, such as those using ResNet-101 as the backbone network and FCN as the object detector. Further research could also be conducted to improve the mAP drop and explore the transferability of UEA to other types of object detection models.

Zheng et al. (2020b) introduced an effective adversarial attack on person re-identification in video surveillance systems known as the dispersion reduction (DR) attack model. This novel approach generates adversarial images that substantially degrade the performance of state-of-the-art person-identification models. Reducing the dispersion of the internal feature map of a neural network makes the objects within the feature map less recognisable and distinguishable, thereby highlighting potential security risks in video surveillance systems. The proposed DR attack model works by iteratively reducing the dispersion of the internal feature map of a neural network to generate adversarial examples. The method updates these adversarial examples on the basis of the gradient of the and keeps the perturbations within a set perturbation budget. The experimental results demonstrated the effectiveness of the DR attack. The adversarial examples generated by this method have a considerable effect on the person's re-identification performance even though the perturbations are imperceptible to the human eye. DR attack outperforms other state-of-the-art attack models, leading to the most notable drop in mAP values.

Li et al. (2022) proposed a novel bidirectional adversarial attack approach against DNN-based object detectors to expose the security vulnerabilities of these models and raise awareness about the potential misuses of object detection techniques. This method involves generating adversarial examples that can mislead the object detection model by minimising the background probability and maximising the ground-truth probability. It introduces a new confidence loss function to decrease foreground scores in RPN. This bidirectional adversarial attack approach consists of two steps: generating adversarial examples and launching adversarial example attacks. The adversarial examples are created using a pre-trained auto-encoder as the generator, and the model is trained with an adversarial approach to improve the similarity between the adversarial examples and the original images and hasten algorithm convergence. These adversarial examples are then employed to attack the object detection model by optimising the perturbations. This attack misleads the model towards the background class and away from the ground-truth class. A confidence loss function is also introduced to reduce the foreground scores in RPN. The simulation results of the proposed bidirectional adversarial attack approach showed that the method achieves a 58.3% mAP drop on the Faster R-CNN object detection model in the white-box attack scenario. In the black-box attack scenario, the adversarial examples generated by the proposed method achieve a 39.5% mAP drop when applied to the YOLOv3 object detection model. These results demonstrate that the proposed bidirectional adversarial attack approach effectively attacks the region confidence and classification accuracy of various object detection models. They prove the strong concealment and satisfactory transferability of the adversarial examples generated by this method, which can achieve substantial interference even in unknown object detection networks with small perturbations that are inconspicuous to humans.

Another reported achievement of this method is a large drop in mAP on object detection models. The adversarial examples generated by this method demonstrate superior transferability because they can also attack different object detection frameworks. Despite the method's effectiveness, the challenge lies in making the generated patches more inconspicuous and less noticeable to humans. Future work could address this issue and further improve the robustness and stealthiness of adversarial attacks against object detectors. Moreover, the attack's effectiveness can still be improved, and possible defences against adversarial attacks on object detection models can be explored.

Thys et al. (2019a) developed an innovative approach called OBJ to generate adversarial patches specifically designed to fool person detectors, thereby highlighting potential vulnerabilities in automated surveillance systems. The objective was to consider a patch that can effectively conceal a person from a person detector, which can be exploited to bypass security systems. The authors deployed an optimisation process to generate these adversarial patches. They focused on minimising three parameters: the non-printability score, the total variation in the image and the objectless score. These patches were trained using actual images of different individuals and applied to these images under various transformations to enhance their robustness. Precision–recall curve analysis demonstrated that the OBJ approach substantially lowered the accuracy of person detection. However, the exact numerical value was not specified in the provided context. In addition, the approach had a low recall percentage, indicating that it is effective in circumventing alarms generated by surveillance systems. Again, the exact numerical value was not specified. The primary challenge encountered in this study was the wide variability in human appearances, different contexts and the absence of a consistent location to place the patch. The authors claimed that future work could concentrate on enhancing the robustness of the patches through additional transformations and improving their transferability across different architectures.

Thakur and Li (2022a) introduced a novel method called pseudo-adversarial training (PAT) designed specifically for detecting adversarial videos that have been manipulated with the intention of deceiving video analysis algorithms. The primary contribution of this study is the development of this new approach, which is geared towards enhancing the robustness of video analysis algorithms against adversarial attacks. The PAT method incorporates a training technique that utilises real and adversarial videos. The process generates adversarial videos through a pseudo-adversarial attack, a modified form of the traditional adversarial attack. These adversarial videos are then used to train the video analysis model, bolstering its resistance to adversarial attacks. The approach also includes a regularisation term to further reinforce the model's robustness. Experiments on UCF-101 and Jester datasets demonstrated that the approach is highly accurate in detecting the adversarial input produced by different attacks. PAT achieves an area under the receiver operating characteristic curve (AUC) of 94.2% on clean and adversarial (containing both sparse and dense attacks) videos. In essence, the experimental results demonstrated that the PAT method outperforms existing methods in detecting adversarial videos. It effectively identifies various types of adversarial videos, including those manipulated to deceive specific video analysis tasks (e.g. object detection and action recognition). One of the key challenges identified in the study is the dynamic nature of adversarial attacks. Attackers can continuously modify their techniques to evade detection, making it challenging to keep pace with their evolving strategies. The authors stated that future research could be directed towards creating advanced detection methods capable of addressing these evolving adversarial attacks. They also suggested the potential application of PAT to areas beyond video analysis, including image classification and NLP.

The proposed research focuses on utilising GAN to launch deceptive attacks by using synthetic images and video samples against AI-powered video surveillance systems. Synthetic samples are generated and used as training data for GAN's generator and discriminator. The contribution lies in introducing an improved GAN model for launching deceptive or fooling attacks against an enhanced YOLO-based object classification system that is embedded in some real-time video surveillance systems. The major factor distinguishing this study from others is that the proposed system is used to determine the extent to which GAN-based fooling attacks can jeopardise ML-augmented video surveillance systems in real time and estimate the failure rate in achieving correct detection of GAN-generated objects (images and videos).

The analysis of related studies provided above raises a question about the efficiency of ML in object detection and malicious behaviour. Therefore, in this study, this concern is addressed to the full extent by evaluating the efficacy of GAN-based attacks in fooling ML-empowered video surveillance systems where a YOLO classifier is used to detect and classify images and videos and decide their authenticity. Several instances of images and video clips can be fed to the proposed solution for training the GAN model and for testing its fooling rate efficiency. These clips include normal scenes and those that portray some malicious human behaviour types, such as humans holding weapons or harmful objects.

Kim and Jung (2017) conducted a thorough examination of illumination-invariant background subtraction methods, a comparison of models currently in use and a discussion of the field's future. They focused on the challenge of background removal in various lighting conditions that can lead to erroneous foreground detection in video surveillance and other applications. The authors evaluated, compared and grouped various models presented in literature into distinct methods to address this problem. Their study highlighted the usefulness of each model in various settings, discussing its advantages and disadvantages. The use of DL algorithms, multimodal data and real-time implementation methodologies were amongst the research trends and possibilities that the authors mentioned for illumination-invariant background removal. This investigation advanced knowledge on background subtraction techniques in the context of light variations.

2.8 Conclusion of Literature Review

This comprehensive literature review has systematically examined the current landscape of object tracking and detection techniques, with particular emphasis on their vulnerabilities to GAN-based adversarial threats. The analysis reveals several critical insights:

First, while deep learning approaches like YOLO, Faster RCNN, and CNNs have demonstrated remarkable performance in object detection tasks, they remain fundamentally vulnerable to sophisticated adversarial attacks. The review has categorized these threats into white-box, black-box, and grey-box attacks, each presenting unique challenges to system security.

Second, Generative Adversarial Neural Networks present a dual-edged sword in this domain. While GANs offer powerful capabilities for generating realistic adversarial examples that can bypass conventional defenses, they also provide promising frameworks for developing robust, tamper-proof systems through adversarial training and data augmentation.

Third, the evaluation of existing defense mechanisms reveals significant gaps in current approaches. Traditional methods like defensive distillation and adversarial training show promise but often lack the adaptability required for real-time video surveillance applications. The computational complexity of many GAN architectures further complicates their practical deployment.

Most critically, the literature demonstrates a pressing need for integrated defense systems that combine the detection capabilities of modern object detectors with enhanced tracking and verification mechanisms. The absence of comprehensive frameworks that address both attack generation and defense evaluation in video surveillance contexts represents a substantial research gap.

This review therefore establishes the foundation for developing an enhanced defensive framework that leverages the strengths of both YOLO-based detection and TLD-based tracking, while incorporating GAN-generated adversaries to create a more robust system. The subsequent chapters will address these identified gaps by proposing and evaluating a fortified YOLO-TLD architecture capable of withstanding sophisticated adversarial attacks in real-time video surveillance environments.

Chapter 3

Methodology and Design

This chapter delves deep into the methodology and design of the research, providing a framework to understand and counter enhanced GAN-generated attacks on AI-empowered video surveillance systems. It starts with a discussion of the research design, emphasising a positivist research paradigm and deductive approach. It includes a comprehensive literature review to inform a proposed model for understanding and mitigating GAN-based attacks and examinations of relevant ML and DL applications in video surveillance. The chapter then presents the modelling process, outlining the role of GANs in generating synthetic attacks and improving surveillance capabilities. Next, it discusses methodology selection, highlighting the need for a customised methodology to address the unique challenges posed by DL applications. The chapter broadens its focus to methodology development, detailing the process of fine-tuning the GAN model across various dimensions and evaluating it against state-of-the-art detection systems. This part includes a comprehensive exploration of transfer learning techniques, training via the gradient descent algorithm, discussion of the loss function and weight update mechanisms and the implementation of the Adam optimiser.

The chapter concludes with an overview of factorisation and its role in enhancing generative models, underscoring its function in dimension reduction for efficient computing and powerful expression within GANs. Overall, the chapter presents a thorough methodology for investigating and strengthening the vulnerabilities of AI-empowered video surveillance systems to GAN-based attacks, providing a foundation for subsequent analyses and insights.

3.1 Research Framework

This research employs a structured, multi-phase experimental framework designed to systematically investigate adversarial attacks and defenses in AI-powered video surveillance systems. The framework consists of five interconnected phases that collectively address the research objectives, as illustrated in Figure 3.1.

3.1.1 Phase 1: Attack Development and Validation

Objective: Develop and validate enhanced GAN-based adversarial attacks (Addressing O1)

- **Input Data:** Standard datasets including COCO, OpenImages, and WIDER Face for training and evaluation
- **GAN Architectures**: Modified DCGAN and cGAN frameworks with enhanced feature learning capabilities.

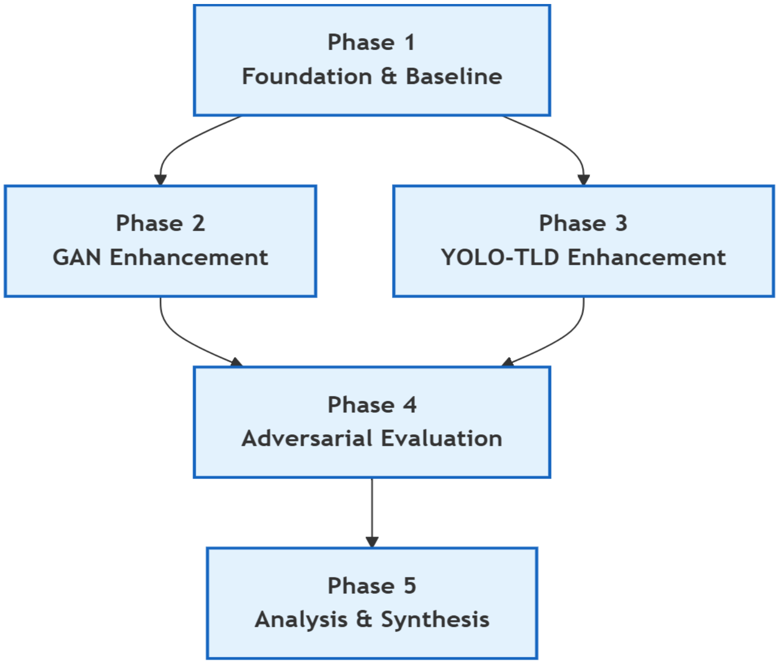


Figure 3.1 Research framework

- **Attack Methods**: Implementation of white-box (FGSM, CW), black-box (scorebased, transfer-based), and grey-box attacks

- **Validation Metrics**: Attack success rate, perturbation magnitude (L2, L∞ norms), and visual quality assessment (SSIM, PSNR)
- **Output:** A repository of high-quality adversarial examples for testing defense mechanisms

3.1.2 Phase 2: Baseline System Vulnerability Assessment

Objective: Evaluate impact of GAN attacks on standard YOLO-based systems (Addressing O2).

- **Baseline Model:** YOLOv4 architecture trained on standard object detection datasets
- **Evaluation Scenarios:**
  - Single-object vs. multi-object detection under attack
  - Performance under varying attack intensities (ϵ values)
  - Real-time processing capability assessment (FPS metrics)
- **Performance Metrics:** mAP (mean Average Precision), precision, recall, F1score degradation under attack
- **Output:** Quantitative analysis of YOLO vulnerability to different attack types

3.1.3 Phase 3: Defense Mechanism Development

Objective: Design and implement fortified YOLO-TLD system (Addressing O3).

- **Architecture Integration:**
  - YOLO detector for initial object detection
  - TLD module for persistent object tracking and verification – Cross-validation mechanism between detection and tracking outputs
- **Defense Components:**
  - Temporal consistency checking across frames
  - Motion pattern analysis for anomaly detection
  - Confidence score fusion from both modules
- **Implementation Details:** Real-time optimization for 30 FPS processing
- **Output:** Integrated YOLO-TLD defense system prototype

3.1.4 Phase 4: Comprehensive Defense Evaluation

Objective: Assess effectiveness of fortified defense system (Addressing O4)

- **Testing Protocol:**
  - Controlled experiments with known adversarial examples
  - Real-world scenario testing with synthetic attacks
  - Stress testing under extreme attack conditions
- **Comparative Analysis:**
  - Performance comparison: Baseline YOLO vs. Fortified YOLO-TLD
  - Computational overhead assessment
  - Robustness across different attack types and intensities
- **Metrics:** Detection accuracy preservation, false positive/negative rates, processing speed maintenance
- **Output:** Comprehensive performance evaluation repor.

3.1.5 Phase 5: Synthesis and Practical Implementation

Objective: Develop practical recommendations and implementation guidelines (Addressing O5)

- **Knowledge Integration**: Combine findings from all experimental phases
- **Guideline Development:** Best practices for robust surveillance system design
- **Implementation Framework:** Deployment considerations for real-world applications
- **Future Research Directions:** Identified gaps and potential extensions
- **Output:** Comprehensive recommendations for industry practitioners and researchers.

3.1.6 Methodological Approach

The research employs a mixed-methods approach combining:

- **Experimental Research:** Controlled laboratory experiments for attack and defense development
- **Quantitative Analysis:** Statistical evaluation of performance metrics under various conditions

- **Comparative Studies:** Systematic comparison between baseline and enhanced systems
- **Empirical Validation:** Real-world scenario testing to validate practical applicability

3.2 Research Design

This research developed an innovative approach to understand and counteract enhanced GAN-based generated attacks on AI-empowered video surveillance systems. A positivist research paradigm, which is traditionally associated with a deductive approach and quantitative methods, was adopted to comprehend and describe the behaviours and vulnerabilities of these surveillance systems.

The approach involved a comprehensive literature review (Chapter 2) to identify the most pertinent studies in the field of AI, DL and GAN-based attacks. This review guided the development of the proposed model for understanding and mitigating GAN-based attacks (Chapter 4). Furthermore, related works involving ML and DL applications in the context of video surveillance were examined to identify the most relevant models for comparison with the proposed approach.

The research methodology for this thesis entailed collecting precise scientific data via measurements from three distinguished datasets: COCO 2020, VOC 2007 and VIRAT. These datasets offer a wide range of image processing and surveillance scenarios and serve as the basis for simulating and analysing GAN-based attacks.

The proposed GAN-based attack model was used to determine the best available approach for detection. Furthermore, the use of three objective evaluation metrics (precision, recall and F1 score) provided evidence of model performance compared with baseline and benchmark models, which are given in Chapter 5.

AI-empowered video surveillance systems against GAN-based attacks under varying conditions were evaluated and tested in this study. The results yielded critical insights into the vulnerabilities of AI-empowered surveillance systems and ways to enhance their resilience against GAN-based attacks.

As previously mentioned, AI-empowered video surveillance systems are highly complex, and their effectiveness depends on the proficiency of their threat detection and response capabilities. A system's readiness against potential real-world threats is tested by simulating attacks using GANs, ultimately bolstering its robustness and efficacy.

The architecture of GANs escalates the adversarial process that is often seen in security systems, with the generator network attempting to create synthetic threats and the discriminator network striving to identify these as non-legitimate. This process, much like the interplay of attack and defence in actual security systems, provides a rich ground for enhancing AI systems' attack detection capabilities.

This research also employed the concept of mental modelling in neuroscience to thoroughly understand the decision-making processes of AI systems. Just as the brain's ventromedial prefrontal cortex and anterior insula play roles in rational thinking and emotional processing, AI systems also engage in similar processes for decision-making under uncertain circumstances, such as identifying potential threats.

This study used a conceptual model in addition to mental modelling. This model helps formalise the mental model's components and provides a framework for analysing and enhancing the AI system's capabilities. These conceptual models aid in simulating realistic threat scenarios, offering a structured approach to developing GAN-based attacks.

Moreover, simulation was conducted to test the AI system's performance against GAN-based attacks. Comparative analyses were conducted with state-of-the-art AI systems and evaluated by using objective metrics, such as MSE, mean absolute percentage error (MAPE) and mean squared logarithmic error (MSLE). State-of-the-art models were chosen as benchmarks because of the success they have demonstrated in the field.

3.2.1 Modelling

AI has been increasingly adopted in various fields, including video surveillance systems. GANs, a key part of this study, are composed of two primary components:

the generator and the discriminator. These components are used to generate attacks against AI-empowered video surveillance systems. This work focused on enhancing GANs' capabilities to generate sophisticated synthetic attacks that can mislead AI-empowered video surveillance systems to determine the possible risk level in the field. After a sample was obtained from latent datasets, the complex process where the generator produces synthetic data that resemble real-world surveillance scenarios or use cases was examined. Concurrently, the discriminator was trained to differentiate between real data and these fake samples, thereby creating a robust surveillance system against GAN-based attacks.

A comparative analysis against current state-of-the-art models was conducted to evaluate the effectiveness of the proposed model. Key performance indicators, such as precision, recall and F1 score, were utilised to ascertain the model's accuracy, sensitivity and overall performance. State-of-the-art models were chosen based on their effectiveness and prevalence in the field at the time of this research.

This work further examined the role of transfer learning techniques, especially under conditions of data scarcity. The aim was to understand how pre-training modules of the proposed model can enhance the generalisation capabilities of GANs in the context of video surveillance. The underlying mechanism of the proposed model is presented in Section 1.2, underpinning the design and parameters of the current experiment. This study underscores the importance of integrating AI with video surveillance systems whilst outlining potential vulnerabilities through GAN-based attacks. It aims to contribute valuable insights to the ongoing discussions on AI security, paving the way for the development of highly resilient video surveillance systems.

3.2.2 Methodology Selection

ML problems often arise between research and deployment, and thorough understanding is required to select an appropriate algorithm and make improvements. Moreover, the development process must be robust and reliable because AI is employed in critical problem domains. Given the broad range of DL applications, no standardised development methodology that can be universally applied has been established. In addition, most ML algorithms are defined to solve a local search problem with an optimised solution. Instead, experts use their knowledge and skills to devise a plan that will likely be developed for certain operations.

The developed solutions for the GAN attack system and the YOLO-defined system consider a local search problem to establish a generalised workflow for ML, as depicted in Figure 3.2. However, these workflows may not be directly applicable to DL where neural networks are utilised, introducing distinct techniques and problem structures at various stages.

The challenges extended beyond conventional ML paradigms. DL models often operate on vast amounts of complex data, requiring specialised techniques for pre-processing, feature extraction and model architecture design. Furthermore, the iterative nature of DL training demands careful attention to hyperparameter tuning and regularisation to prevent overfitting. Deployment considerations for DL systems also differ, with concerns (e.g. model optimisation) for inference speed and resource constraints becoming paramount. Despite these complexities, efforts are underway to establish standardised workflows for DL development, aiming to streamline processes and ensure reproducibility. However, the dynamic nature of DL research and the rapid evolution of techniques mean that such workflows have to remain flexible and adaptable to accommodate emerging methodologies and technologies.

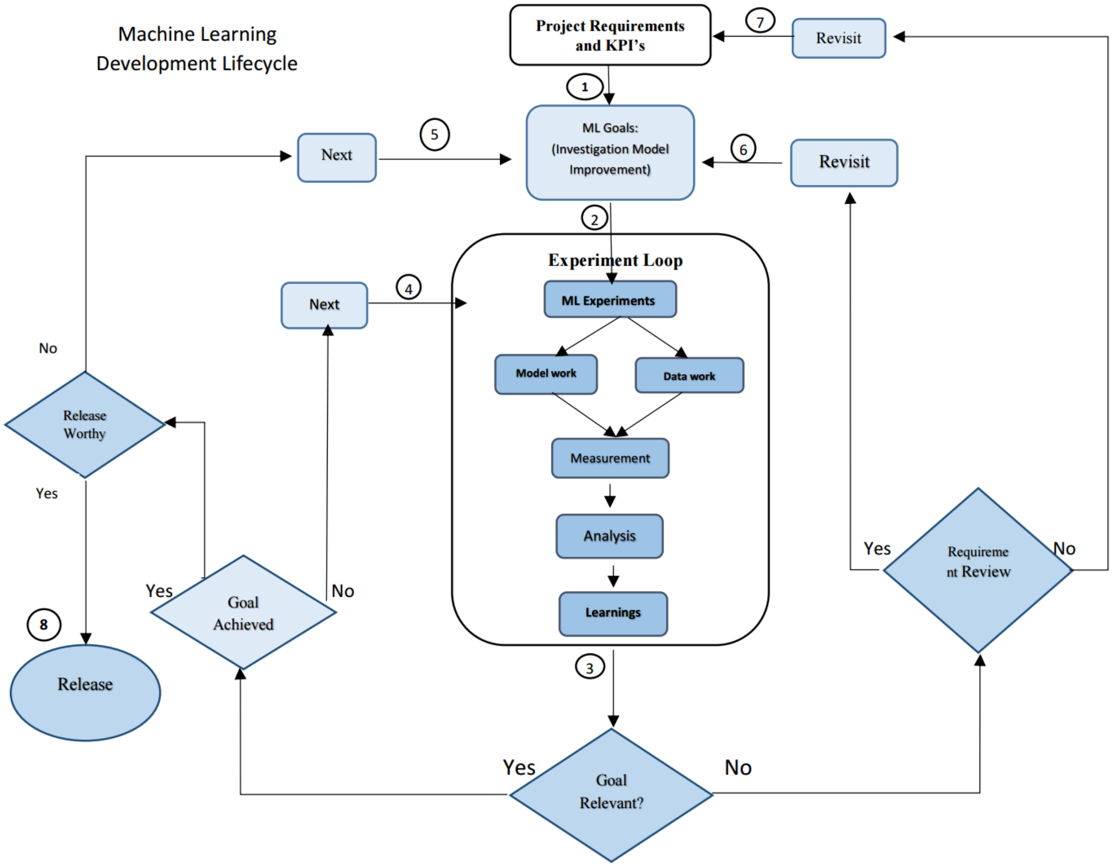


Figure 3.2 Draft of the machine learning development life cycle (Pinhasi, 2021)

3.3 Design Architecture

The proposed fortified object detection system integrates three core components: a YOLO-based detector, a TLD tracking module, and GAN-based adversarial training. The architecture is designed to operate in real-time while maintaining robustness against adversarial attacks. The architecture is illustrated in Figure 3.3.

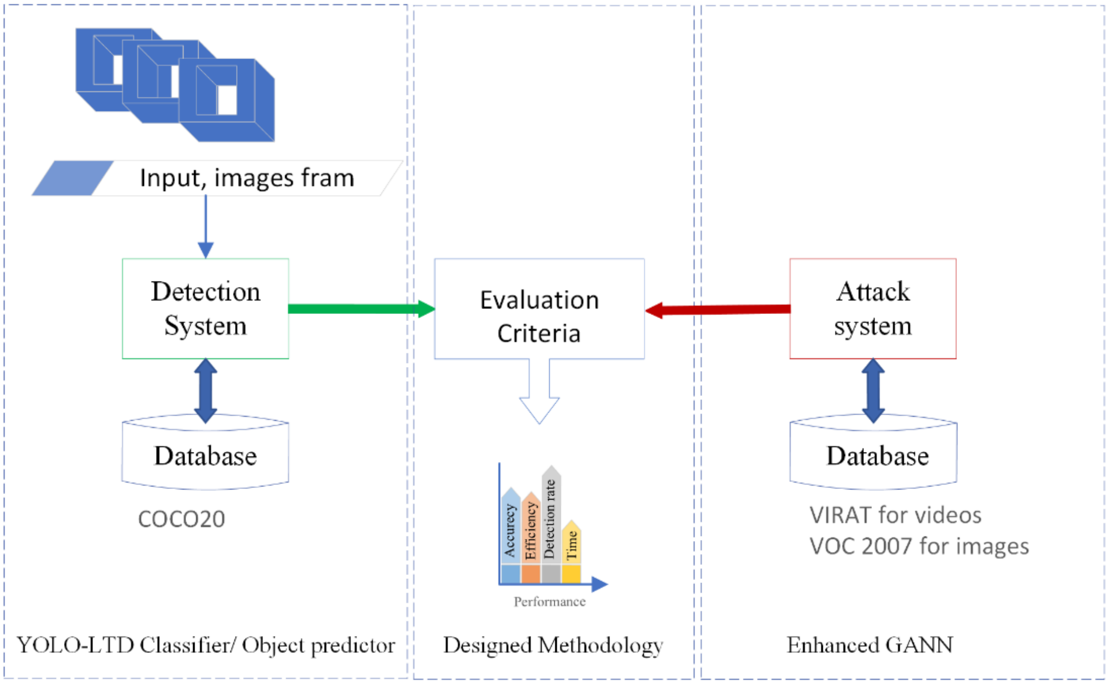


Figure 3.3 Design architecture

3.3.1 Component Specifications

YOLO Detection Module

- **Architecture:** YOLOv4 with DarkNet-53 backbone
- **Input Resolution:** 416 × 416 pixels
- **Output:** Bounding boxes, class probabilities, confidence scores
- **Multi-scale Detection:** Three detection scales (52 × 52, 26 × 26, 13 × 13)
- **Real-time Performance:** Target 25-30 FPS processing

TLD Tracking Module

- **Components:**
  - **Tracker:** Median Flow tracker for short-term object tracking – **Detector:** Random ferns classifier for object re-detection
  - **Learner:** Online learning for model updates and adaptation
- **Features:** Optical flow, texture patterns, motion consistency
- **Temporal Window:** 10-frame history for trajectory analysis

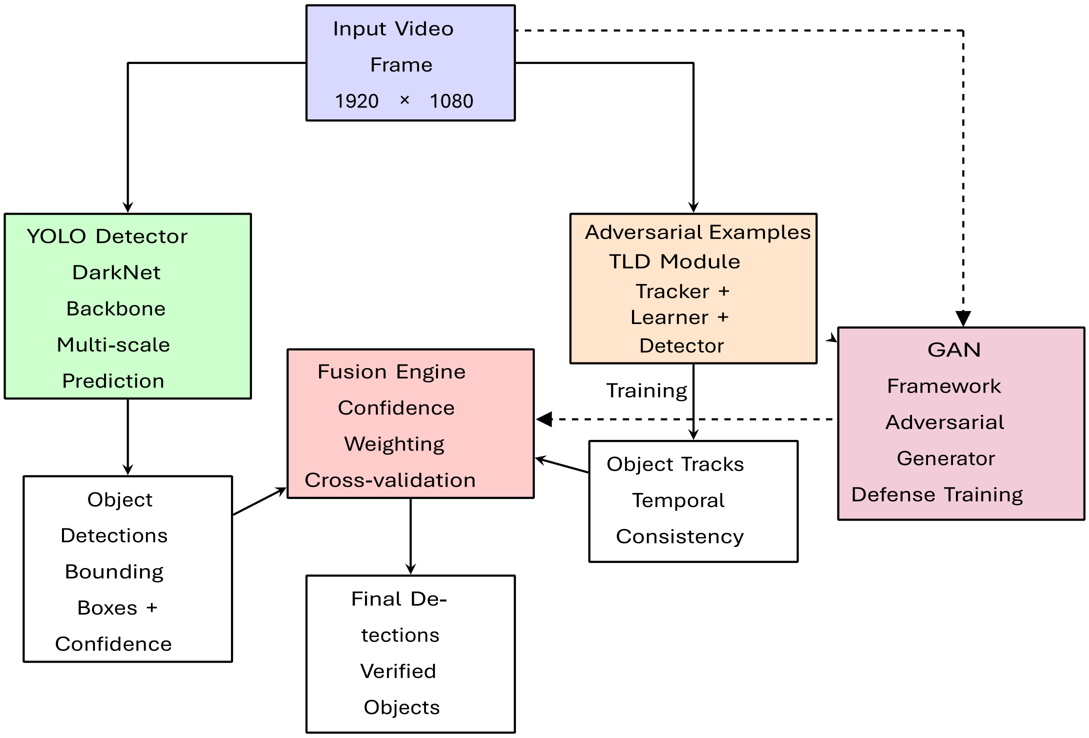


Figure 3.4 Detailed Architecture of Fortified YOLO-TLD System

Fusion Engine

- **Confidence Fusion:** Weighted combination of detection and tracking confidence scores
- **Cross-validation:** Spatial and temporal consistency checks between modules
- **Decision Logic:**
  - High confidence in both: Accept detection
  - Conflict between modules: Temporal analysis and trajectory prediction
  - Low confidence: Trigger re-detection or mark as uncertain

GAN Training Framework

- **Generator:** DCGAN architecture for adversarial example generation
- **Attacks Simulated:** FGSM, PGD, CW attacks with $\epsilon \in [0.01, 0.3]$
- **Training Data:** COCO dataset augmented with adversarial examples
- **Defense Strategy:** Adversarial training with generated attacks

3.3.2 Data Flow Process

1. **Frame Input:** HD video frame (1920 × 1080) is received

2. **Parallel Processing:** Frame is simultaneously processed by YOLO detector and TLD tracker
3. **Detection:** YOLO identifies objects and provides initial bounding boxes
4. **Tracking:** TLD maintains object trajectories and provides temporal consistency
5. **Fusion:** Fusion engine combines outputs using confidence-weighted averaging
6. **Validation:** Cross-module verification and temporal consistency checks
7. **Output:** Final verified detections with enhanced robustness metrics

3.3.3 Performance Targets

- **Processing Speed:** ≥ 25 FPS for real-time operation
- **Detection Accuracy:** mAP ≥ 75% on clean data
- **Robustness:** ≤ 15% performance degradation under adversarial attacks
- **Precision:** ≥ 80% in challenging conditions (occlusion, scale variance)

3.4 Methodology Development

GANs are well-suited for DL tasks because of their powerful encoding capacity; using CNNs as its core architecture, a GAN can learn features from high layers with few parameters. Thus, implementing the mathematics behind GANs provided invaluable insights into how these intelligent applications work. The goal is to train a generator network to produce samples that are similar to a target distribution, and a discriminator network tries to distinguish the generated samples from the real ones. The research was conducted in three stages.

Stage 1: State-of-the-Art Detection System

Pre-processing focused on data preparation and building the knowledge base for the ML algorithm. Relevant datasets were prepared from different sources, such as video surveillance clips from open-source resources. Additional open-source datasets, such as COCO and VOC for images and VIRAT for videos, were also used. This stage included data collection, filtering and analysis. The data were divided into two: training and testing sets. The results of the data analysis determined the necessary parameters to detect malicious behaviour (attacked images/videos using GAN) patterns via video surveillance.

A literature review of state-of-the-art techniques used for object detection was conducted, and the best techniques were identified by implementing them (in terms of detection accuracy, time-related performance, etc.).

The YOLO technique was determined to be the best classifier amongst the ones that were investigated. Its performance was measured on static images (the COCO dataset) and video data (e.g. VIRAT). Samples of images were used to detect interesting objects, such as humans and weapons, in the original and attached images. This stage included adopting the best object detection algorithms and adaptation with TLD object tracking. This combination of object detection by YOLO and tracking by TLD enhanced the detection rate for static images and video. The system collected, trained and deployed the new proposed detector to be used for the next stage.

Object detection is an essential task for this research, and it should have high accuracy and high-speed performance because it is applied to video analysis, including adopting the best object detection algorithm with object tracking by using object ID. Challenges, such as deformation, scale changes, illumination variations and partial occlusion, often hinder traditional object-tracking methods. TLD is a well-established and effective long-term tracking algorithm capable of addressing these issues (Liu et al., 2023; Sreeja et al., 2023). Its real-time performance becomes crucial when such algorithms are implemented in real-world scenarios. Thus, enhancing the performance of YOLO detectors by adding an object tracking algorithm based on TLD improves the defence system. The COCO 2020 dataset was used in this study, and a new algorithm was developed for the adapted neural network tracking system to detect the object, as shown in 3.3. The system collected, trained and deployed the object tracking to be used for the next stage. The detailed steps conducted at this stage are as follows:

- *Initialise the TLD object tracking algorithm:* The TLD object tracking algorithm is initialised by using the initial position and size of the object that needs to be tracked.
- *Detect objects in the current frame:* OpenCV and YOLO are used to detect objects in the current frame of the video. YOLO updates the object’s position and size in the TLD algorithm.

- *Update the object's appearance model:* The TLD learning algorithm is used to update the object's appearance model on the basis of the detected objects in the current frame. This process improves tracking accuracy over the processing time.
- *Track the object in subsequent frames*: The TLD tracking algorithm is employed to keep track of the object as it moves through the subsequent frames of the video. The detected objects in each frame are updated by the object's position and size as needed.
- *Repeat the process for multiple objects:* The process is repeated to track multiple objects for each object, and separate instances of the TLD object tracking algorithm are used for each object.

Implementing techniques such as Kalman filtering (Cai et al., 2023) or particle filtering alongside TLD can refine the tracking predictions and enhance the robustness and reliability of the tracking system, especially in scenarios with occlusions or abrupt object motions. Moreover, incorporating feature-based tracking methods, such as KLT or SIFT, can further bolster tracking accuracy by leveraging distinctive object features. Furthermore, integrating DL-based re-identification models aids in handling object identity reappearances, ensuring consistent tracking across frames. These enhancements collectively fortify the tracking system's performance, making it adept at managing diverse real-world scenarios with multiple moving objects.

Stage 2: Enhanced GAN Model

This stage focused on the enhancements made to the GAN method to make it produce highly deceptive (close to original) fake images and videos. These images and videos are then used to fool YOLO-based video surveillance systems. Object detection is an essential task in this research because it is used to evaluate the risk of using GANs as an ethical attack vector (for testing purposes). Given that it is applied to video analysis, object detection must be implemented with high accuracy and high speed.

For this reason, enhancing the loss function used by GAN is a key factor in improving the generation of highly faked, deceptive images and videos. In this study, enhancement of the GAN method was conducted by producing four different variants. These variants were obtained by modifying the image pixel size and through a newly developed and proposed loss function. The four variants were used to assess the effect of the new loss function on the GAN-based attack system's cooling capability.

Adversarial attacks in ML refer to instances where an attacker intentionally manipulates input data to cause an ML model to make incorrect or harmful predictions. These attacks can be particularly concerning when the model is used for decision-making, specifically in surveillance systems.

GANs are well-suited for DL tasks because of their powerful capacity for encodings; using CNNs as its core architecture, a GAN can learn features from high layers with few parameters. Thus, implementing the mathematics behind GANs provides invaluable insights into how this intelligent application operates in actual operating systems.

The GAN technique developed in this work aims to train a generator network to produce samples similar to a target distribution. Then, a discriminator network distinguishes the generated samples from the real ones. Image and video processing encounter different types of adversarial attacks, including the following (Chakraborty et al., 2018; Sun et al., 2022):

1. Evasion attacks: These attacks involve manipulating the input data to cause the model to misclassify them. For example, an attacker might add noise to an image to cause a model to incorrectly classify it as a different object.
2. Poisoning attacks: These attacks involve modifying the training data to cause the model to behave poorly on certain inputs. For example, an attacker might add a small number of maliciously crafted training examples to a dataset to cause the model to misclassify a specific input type.
3. Trojan attacks: These attacks involve adding a 'backdoor' to the model that allows the attacker to manipulate the model's behaviour in a specific way. For example, an attacker might add a small, almost imperceptible change to the input data that causes the model to behave differently.

4. Overfitting attacks: These attacks involve training a model on a dataset specifically designed to cause it to overfit or perform well on the training data but poorly on new, unseen data.
5. Model inversion attacks: These attacks involve using a model's output to infer sensitive information about the input data. For example, an attacker might use a model that predicts medical diagnoses to infer sensitive information about a patient's medical history.

This research focused on the evasion attack because it creates adversarial examples that evade detection or classification by the GAN-based system. Attackers can exploit vulnerabilities and generate fake images or videos that appear legitimate to the model but may contain subtle modifications designed to trigger incorrect responses. These evasion techniques are commonly used to test the robustness of GAN models in video surveillance systems, highlighting the importance of implementing effective defence mechanisms to counter such adversarial manipulation and ensure the system's security and reliability.

One common approach that was considered for this research is to add noise or perturbations to the input data that are small enough to be imperceptible to humans but cause the model to misclassify the data. These perturbations can be added to images through various methods, such as gradient descent or evolutionary algorithms.

Another approach is to use transfer attacks, in which the attacker creates adversarial examples that can fool a specific model but are ineffective against other models. This can be done by training a model to recognise the adversarial examples and using that model to generate examples specifically designed to fool the target model.

Stage 3: Performance Evaluation

The newly generated images are applied again with the YOLO classifier to verify the accuracy of object detection. The evaluation criteria are based on the variance between the number of objects in the test sample images and the correct detection of objects within the same image. Additionally, efficiency decreases after the GAN attack is calculated based on the disparities between the detected images before and after the attack; the YOLO classifier serves as the baseline detection system. The fooling rate efficiency and efficiency drop are determined in accordance with the following equations:

$$\text{Efficiency of fooling rate} = \frac{(\text{No. of test Objects}_{\text{Original image}} - \text{Correct object detected}_{\text{After attack}})}{\text{No. of test Objects}_{\text{Before attack}}} \quad , \qquad (3.1)$$

where the numerator is the total number of test objects (or images) that are correctly detected by the model before the attack and the denominator is the number of originally correct detections that are now incorrect (or missed) after the attack is applied.

$$\text{Efficiencydrop} = \frac{(\text{ Correctly detected by YOLO classifier}_{\text{before attack}} - \text{Correct object detected by YOLO classifier}_{\text{After GAN attack}})}{\text{Correctly detected by YOLO classifier}_{\text{Before attack}}} \qquad (3.2)$$

3.4.1 Training Process

The training process relied on the gradient descent (GD) algorithm to update the model parameters simultaneously during each step. At this stage, two mini-batches were sampled (one with x values from the dataset and another with z values taken from the model’s prior over latent variables) as described by the pseudocode in Figure 3.5.

```
Start GAN training
Identify the data set, number of epochs and number of
batches:
    calculate the number of batches per epoch =
    int(len(dataset) / number of batches)
    calculate the number of training iterations = number
    of batches per epoch * number of epochs

for i in range(number of training iterations):
    update the discriminator model
    update the generator model
```

Figure 3.5 GAN training algorithm pseudocode

The training dataset was split into two batches for each epoch. The first batch was used as an input to the generator model for generating fake samples (to generate controlled noise). The second batch was adopted as an input to the discriminator as a real sample. The discriminator model classified the generated and real images. Then, the weights were updated accordingly. After that, the generator model used a batch of random points from the latent space to generate additional fake images and then sent these images to the discriminator for classification. The feedback was used to update the weights of the generator model. The whole process is explained in the pseudocode in Figure 3.6.

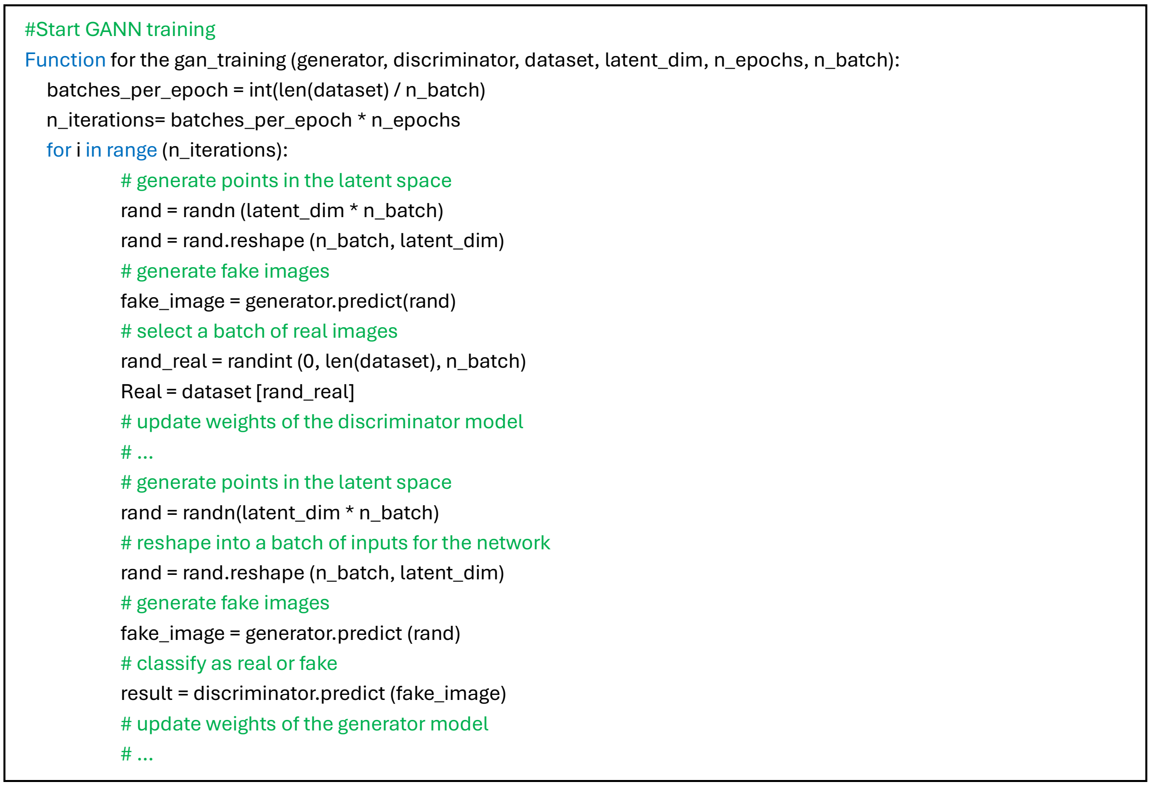
Notably, the generator is updated with only a single batch of samples each iteration, whereas the discriminator is updated with two batches of samples.

3.4.2 Loss Function and Weight Update

Minibatch stochastic CosGAN was applied. The process involved training GAN on small batches of data rather than the entire dataset to improve the efficiency and speed of training whilst helping to prevent overfitting.

Minibatch stochastic CosGAN training of GANs is explained as follows:

1. Initialise the generator and discriminator networks with random weights and biases.
2. Divide the training dataset into small batches of size B.
3. For each minibatch:
   a) Generate a set of synthetic data samples using the generator network.
   b) Concatenate the synthetic data samples with a set of real data samples from the minibatch.
   c) Train the discriminator on this combined set of data by using CosGAN and the cross-entropy loss function.
   d) Train the generator by using the output of the discriminator as feedback and with CosGAN and the cross-entropy loss function.
4. Steps c and d are repeated until the model converges, or a predetermined number of epochs is reached.



```
#Start GANN training
Function for the gan_training (generator, discriminator, dataset, latent_dim, n_epochs, n_batch):
  batches_per_epoch = int(len(dataset) / n_batch)
  n_iterations= batches_per_epoch * n_epochs
  for i in range (n_iterations):
        # generate points in the latent space
        rand = randn (latent_dim * n_batch)
        rand = rand.reshape (n_batch, latent_dim)
        # generate fake images
        fake_image = generator.predict(rand)
        # select a batch of real images
        rand_real = randint (0, len(dataset), n_batch)
        Real = dataset [rand_real]
        # update weights of the discriminator model
        # ...
        # generate points in the latent space
        rand = randn(latent_dim * n_batch)
        # reshape into a batch of inputs for the network
        rand = rand.reshape (n_batch, latent_dim)
        # generate fake images
        fake_image = generator.predict (rand)
        # classify as real or fake
        result = discriminator.predict (fake_image)
        # update weights of the generator model
        # ...
```

Figure 3.6 GAN training algorithm pseudocode with preparation for updating model weights

3.4.3 Adam Optimiser

The Adam optimiser was utilised to adjust the network's weights and biases to minimise the loss function. This optimiser, a variant of stochastic GD, employs adaptive learning rates by adjusting the second-order rate using historical gradient data. Extending beyond the basic GD algorithm, Adam incorporates the first- and second-order moments of the gradients, enabling rapid adaptation to the loss function's curvature. Widely favoured in deep learning, Adam often delivers excellent performance with minimal hyperparameter fine-tuning.

When the Adam optimiser is utilised in a GAN, it is commonly designated as the optimiser during model definition and subsequently passed to the 'fit' function for training. Its implementation in TensorFlow 2 is shown in Figure 3.7.

```
optimizer = tf.keras.optimizers.Adam()
model.compile(optimizer=optimizer, loss='binary_crossentropy')
model.fit(x_train, y_train, epochs=10)
```

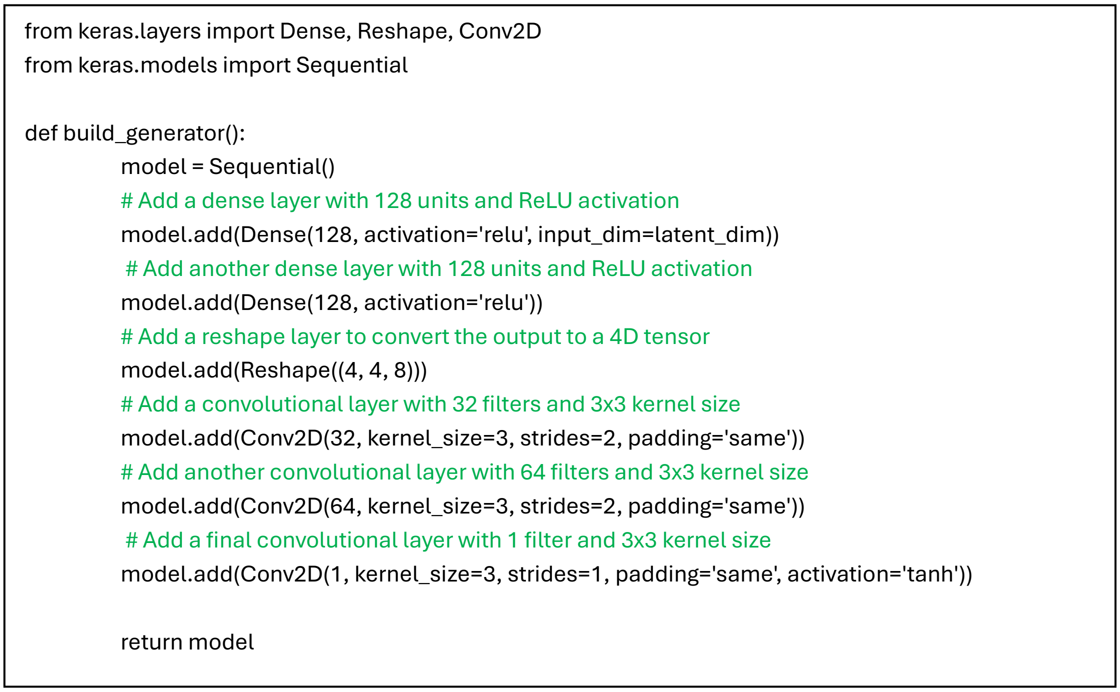
Figure 3.7 Adam optimiser demonstration using TensorFlow 2

For the GAN model in Figure 3.7, which is presented for ‘xtrain’ and ‘ytrain’, the input and output data were used for training, and ‘epochs’ refers to the number of training iterations. The Adam optimiser was used to update the model’s weights and biases to minimise the CosGAN loss function, which in this case is binary cross-entropy.

3.5 GAN Implementation

The whole process is demonstrated in pseudocode, as shown in Figure 3.8. This code defines a function called ‘build_generator’ that creates the generator network via the Keras library. The generator network consists of a series of dense and convolutional layers, with ReLU activation functions for the dense layers and a tanh activation function for the final convolutional layer. The input to the generator is a noise sample from the latent space, and the output is a synthetic image.

When this generator network is used, the latent space dimension is defined as ‘*latent_dim*’, and the ‘*build_generator*’ function is called to create the model. Then, it is compiled and trained using the Keras ‘*compile’* and ‘*fit’* functions.

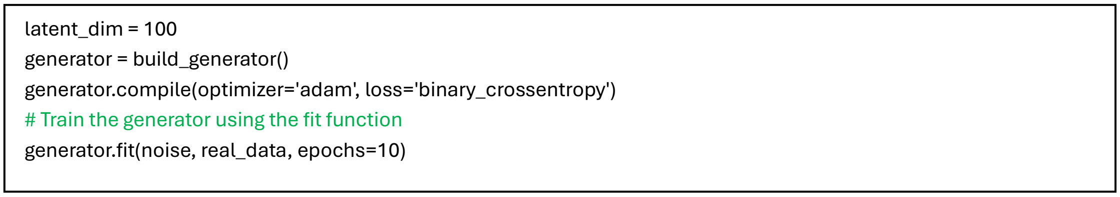


```
from keras.layers import Dense, Reshape, Conv2D
from keras.models import Sequential

def build_generator():
        model = Sequential()
        # Add a dense layer with 128 units and ReLU activation
        model.add(Dense(128, activation='relu', input_dim=latent_dim))
         # Add another dense layer with 128 units and ReLU activation
        model.add(Dense(128, activation='relu'))
        # Add a reshape layer to convert the output to a 4D tensor
        model.add(Reshape((4, 4, 8)))
        # Add a convolutional layer with 32 filters and 3x3 kernel size
        model.add(Conv2D(32, kernel_size=3, strides=2, padding='same'))
        # Add another convolutional layer with 64 filters and 3x3 kernel size
        model.add(Conv2D(64, kernel_size=3, strides=2, padding='same'))
         # Add a final convolutional layer with 1 filter and 3x3 kernel size
        model.add(Conv2D(1, kernel_size=3, strides=1, padding='same', activation='tanh'))

        return model
```

Figure 3.8 Pseudocode of GAN modelling using Keras



```
latent_dim = 100
generator = build_generator()
generator.compile(optimizer='adam', loss='binary_crossentropy')
# Train the generator using the fit function
generator.fit(noise, real_data, epochs=10)
```

Figure 3.9 Pseudocode of the GAN training parameters using Keras

As shown in Figure 3.9, ‘*noise’* is a tensor containing noise samples from the latent space, and ‘*real_data’* is a tensor containing real data samples from the training dataset. When the Adam optimiser is used, the learning rate needs to be specified; this rate determines the size of the optimiser’s step in the direction of the gradient. Moreover, other hyperparameters, such as the decay rates for the first and second moment estimates (beta1 and beta2) and the tolerance for the relative change in the loss function (epsilon), are adopted. The term ‘*model’* refers to a Keras model that is being trained to minimise the MSE loss using the Adam optimiser, as shown in Figure 3.10. The learning rate is set to 0.001, and the model is trained for 10 epochs by using the '*fit*' function.

```
from keras.optimizers import Adam
# Define the Adam optimizer with a learning rate of 0.001
optimizer = Adam(lr=0.001)
# Compile the model with the Adam optimizer and mean squared error loss
model.compile(optimizer=optimizer, loss='MSE')

# Train the model using the fit function
model.fit(x_train, y_train, epochs=10)
```

Figure 3.10 Pseudocode of Adam optimiser modelling using Keras

Factorisation

CNNs scan an image within discrete portions that can be manipulated independently to obtain features. With GAN, an additional factorising element is provided by dimension reduction. This added element often improves the accuracy of generative models by allowing for an efficient representation of information. Computationally efficient models are created by factorising convolutions. Much of the efficiency and effectiveness of expression obtained using generative models stems from the generous use of dimension reduction in Google networks. Therefore, using GANs with CNNs yields highly expressive local representations, making generative models highly effective and accurate. By leveraging the generative capabilities of GANs, combined with the foundational power of CNNs, we obtain generative models that are accurate and computationally efficient.

Dimension reduction enables even successful generative modelling whilst simultaneously mitigating certain computational challenges. It also promotes popular GAN applications, such as CGANs, to deep levels and provides ways to overcome other deep learning situations where traditional methods lack efficiency. In short, generative modelling enabled by dimension reduction is advantageous for efficient computing and powerful expression needs. Furthermore, with suitable factorisation, it results in disentangled parameters that can be trained fast even when only a single computer is employed. Such an approach provides computational and memory savings, enabling us to develop large filter-bank sizes whilst still using a single computer for training because of the needed parameter reduction. This means an effective optimisation process can improve performance when applied to networks producing high-resolution images. For example, a 5×5 convolution with *n* filters over a grid with *m* filters is 25/9 = 2.78 times more computationally expensive than a 3x3 convolution with the same number of filters, as shown in Figure 3.11. Therefore, a small, fully connected neural network sliding over 5x5 tiles over its input suggests that multiple small layers may be even more effective than one large layer alone.

Large network architectures may still yield superior results, and they could represent a breakthrough in condensing network complexity with little loss in encoder efficiency. Thus, this method appears promising for designers with limited computing power or those exploring new design ideas that are unsupported by traditional CNNs.

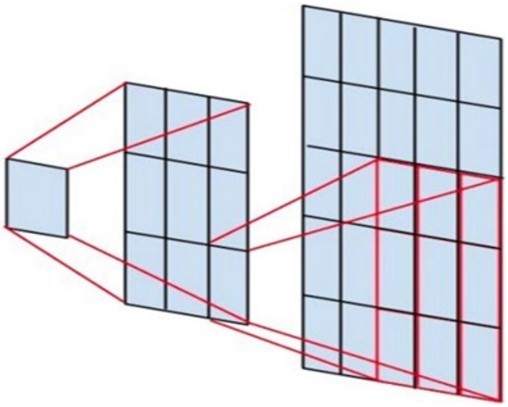


Figure 3.11 Image factorisation in CNN processing

When synthetic images are generated with enhanced GAN modelling, the generator is fed random noise as an input, which it uses to generate an image. By manipulating the noise input, certain aspects of the generated image, such as the colour, shape, or texture of the object depicted in the image, can be manipulated. An ML model can be fooled by generating synthetic images that are designed to be difficult for the model to classify correctly.

An adversarial attack embeds a noise vector in the input image. This noise vector is intentionally designed to confuse the AI-based object detection system with DL models. The pixels in the noise vector are “equal to the sign of the elements of the gradient of the cost function concerning the input image” (Goodfellow et al., 2014b).

3.6 Technical Details

3.6.1 Programming Language

Selecting an appropriate programming language is crucial for successful development and progress in DL. Gathering insights from experts in this field is valuable in determining the ideal language for a specific field. Python is the preferred choice in ML and DNNs (Raschka et al., 2020).

Python’s popularity can be attributed to its numerous advantages for ML developers, including a vast library ecosystem, platform independence and a large supportive community. Although some developers contemplate using R for ML and DL, it is not widely adopted in the field. Only 31% of developers use R, whereas 87% employ Python (Hayes, 2020). This discrepancy aligns with R’s intended purpose, which caters to data analysts and scientists with mathematical backgrounds rather than those with programming or computer science backgrounds (r-project, n.d.). Consequently, R may not be well-suited for DL because it falls outside the scope of its design.

A suitable alternative for developers is C/C++, but it remains a distant second to Python. Although C++ is more attractive for computationally intensive tasks such as DL, Python offers the advantages of ease of learning and practical usage (Tsai et al., 2021). Moreover, most published studies on ML employed Python as the language of choice for implementing models (Kanungo, 2023), giving Python a competitive edge. Therefore, for this project, the chosen programming language is Python, a highly popular language with extensive support for DL and a vast community that can assist.

Opting for Python enables the developer to easily implement and build upon existing work because most of the models described in literature were likely developed using Python.

3.6.2 Libraries

As mentioned earlier, Python offers an extensive library collection that provides the developer with ready-to-use code for each phase of the development process. The primary library selected for this project is TensorFlow, a robust library specifically designed for generative DL. TensorFlow utilises the Keras library as an API endpoint to facilitate ML tasks. Other libraries can be employed for different stages of development. For instance, OpenCV can be utilised for data pre-processing, and the pre-processed data can be stored in a Panda data frame for analysis. Moreover, the NumPy library demonstrates a faster, more efficient performance compared with the built-in list data type in Python, making it a preferred choice for managing arrays.

In summary, Python’s rich library ecosystem allows developers to use pre-existing code for every aspect of their project. For this particular project, TensorFlow is the main library used, with support from additional libraries, such as OpenCV for pre-processing and Panda for data analysis. NumPy arrays from the NumPy library are also used to enhance efficiency and speed during implementation.

3.6.3 Integrated Development Environment

Instead of opting for an integrated development environment (IDE), this project utilises Visual Studio Code (VSCode), which is a code editor. VSCode has gained substantial popularity and is regarded as the most used code editor according to the 2021 Stack Overflow Developer Survey (Bhatia, 2021). According to the survey, 71% of developers consider VSCode their primary code editor choice. VSCode has gained immense popularity amongst developers because of its lightweight and adaptable architecture, which consolidates various functionalities, including version control, within a single editor. It offers many community-developed and supported extensions that considerably enhance development efficiency. These extensions provide features, such as code snippets, syntax highlighting and IntelliSense, that aid in coding tasks. Notably, Python’s supporting extensions are amongst the most widely installed extensions, highlighting the popularity of Python amongst VSCode users.

3.6.4 Hardware

The hardware specifications for this project are specifically tailored to meet the demands of DL, which is the project’s core focus and requires substantial computational power. The primary consideration is selecting a suitable GPU. GPUs are preferred and widely utilised in DL because of their superior ability to manage concurrent computations compared with central processing units (CPUs). The following criteria are considered when choosing a GPU for DL:

- Memory bandwidth
- Number of cores
- Processing power
- Video RAM size

For DL, a GPU of at least 8 GB is recommended. Ideally, the 3080, 3080ti and 3090 GPUs from NVIDIA are suitable choices. These GPUs offer AI-enabled functionalities and support popular deep-learning frameworks (Mungoli, 2023).

Another important consideration is the system’s random-access memory (RAM). A large RAM capacity is recommended for large datasets. Whilst DL models can be trained with a minimum of 8 GB of RAM, 16 GB of RAM or higher is prescribed for image classification solutions (Mijwil et al., 2023). The CPU is the final factor to consider. An Intel Core i7 processor is the minimum requirement for ensuring consistent performance. CPUs with equivalent or higher processing power should be sufficient.

Given that these demanding hardware requirements are not typically available on standard consumer-grade computers, two viable options for meeting these requirements are provided as follows:

1) Utilising Google’s hardware through Google Colab notebooks. A workspace from Amazon, which provides a virtual machine with the necessary hardware specifications, can be rented from any device.

2) A local PC for multiple training and testing of GAN. A local PC with 16 GB of RAM and an Intel Core i7 processor with 8 GB of GPU and 3060 NVIDIA can be adopted.

By leveraging these alternatives, the required hardware specifications for DL can be met effectively.

3.7 Summary

The research design paradigm adopted in this study revealed that YOLO-based systems are the best classifiers that can be used to detect synthetic images and videos; hence, these systems were selected as the enhancer of the video surveillance system. In addition, exploring related studies involving ML and DL applications in the context of video surveillance helped identify the most relevant models for comparison with the proposed approach.

Furthermore, after careful consideration, three objective evaluation metrics (i.e. precision, recall and F1 score) were employed to examine model performance and compare the proposed model with baseline and benchmark models.

Firstly, the methodology was structured around building an understanding of how GANs can model AI-generated attacks on video surveillance systems, hence answering RQ1. Secondly, studying the advantages of incorporating Monte Carlo dropout and orthogonal kernel weight initialisation techniques within the GAN model enhanced the system's detection capability, thus addressing RQ2. The methodology's third aspect lies in applying partial transfer learning techniques to recommend practical solutions for RQ3, which relates to improving the reliability and robustness of AI-empowered surveillance systems against GAN-based attacks. Towards the end, the evasion attack model, which involves manipulating input data to cause the model to misclassify, was adopted as an attack model during the experimentation.

Chapter 4

Proposed Framework: Enhanced Defence and Attack Models

4.1 Chapter Overview

This chapter presents the core technical contributions of this work, addressing Research Questions 2 and 3 and fulfilling Research Objective 3. It is structured into two main parts.

1. Defence System: A novel framework that fortifies the YOLO object detector by integrating it with a TLD module to enhance its robustness against adversarial attacks.
2. Attack Model: An enhanced generative adversarial neural network is adopted.

The following sections present the architecture, theoretical foundations and algorithmic formulations of both components.

4.2 YOLO-TLD Defence Framework

This section examines the proposed YOLO-TLD framework that integrates the YOLO object detector with the TLD paradigm to create a robust system that is resilient to occlusions and adversarial perturbations.

4.2.1 System Architecture and Theoretical Foundation

The core innovation of the YOLO-TLD framework is the synergistic fusion of YOLO's high-accuracy detection with TLD's long-term tracking robustness. This integration overcomes the individual limitations of each component: YOLO's susceptibility to losing track identity through occlusions and the classic TLD's reliance on a weak detector. The architecture, depicted in Figure 4.1, operates through three interconnected components.

1. YOLO detector: Acts as a high-recall, class-specific proposer that performs a full-frame scan in every frame to provide candidate detections independent of the tracker's state.

2. Tracker: Utilises a short-term tracking module (e.g., median flow) to estimate the object's motion between consecutive frames. It is efficient but fails under rapid motion or full occlusion.
3. Learner (P-N learning): It is the core adaptation module. It employs an online learning mechanism to continuously update a patch-based classifier.
   - P-Expert: Identifies new positive examples when the tracker and YOLO detector agree, reinforcing stable tracks.
   - N-Expert: Identifies negative examples from erroneous detections or background patches, suppressing false positives.

The integration is managed by a cascade detector, which combines the outputs of all three components. A final detection is accepted only after passing through this cascade, which validates spatial consistency, temporal coherence and the online learner's confidence.

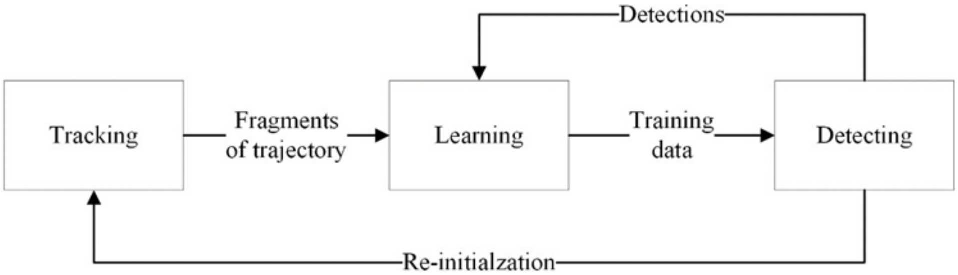


Figure 4.1 System architecture of the proposed YOLO-TLD framework

4.2.2 YOLO-TLD Algorithm

The step-by-step procedure for the YOLO-TLD framework is formalised in Algorithm 1. This algorithmic formulation ensures robustness to tracker drift (via YOLO), resistance to occlusions (via re-acquisition) and adaptability to appearance changes (via online learning). Parameters τ_{agree} and τ_{conf} control the agreement and confidence thresholds, respectively.

4.3 Enhanced GAN Attack Model

An enhanced GAN attack model was developed to rigorously evaluate the proposed YOLO-TLD defence. This section presents its detailed architecture and the novel loss function that forms the core of its improvement.

4.3.1 GAN Architecture and Training

GAN adopts a standard framework comprising a generator (*G*) and a discriminator (*D*) engaged in a two-player minimax game. The generator learns to map a random noise vector z to a synthetic sample G(z), and the discriminator learns to distinguish between real samples (x) and generated samples $G(z)$. The training process involves alternating updates to *D* and *G* to optimise the adversarial objective.

Algorithm 4.1 YOLO-TLD for robust object tracking

```
Require: Video Stream $V$, Target Object Bounding Box $B_0$ in initial frame $I_0$, Target Class $C$
Ensure: Sequence of bounding boxes $\{B_t\}$ for the tracked object in each frame $I_t$
Initialisation:
        Load pre-trained YOLO weights for class $C$
        Initialise P-N Learner with image patch from $B_0$ in $I_0$
        Initialise Median Flow tracker with $B_0$ on $I_0$
        $Bprev \leftarrow B0$

for each frame $I_t$ in $V$ do
        Parallel Execution:
            $B_{track} \leftarrow$ Tracker.predict($I_t$, $B_{prev}$)
            $D_{yolo} \leftarrow$ YOLO.detect($I_t$, $C$)                ▷ Set of detections $\{d_i\}$

        Cascade Detection & Integration:
            Apply spatial variance filter to $D_{yolo}$
            For each $d_i$ in filtered $D_{yolo}$, get learner confidence $s_i$

            Case 1 (Tracker & YOLO Agree):
                If IoU($B_{track}$, $d_i$) > $\tau_{agree}$ and $s_i > \tau_{conf}$:
                    $B_t \leftarrow$ weightedAverage($B_{track}$, $d_i$)
                    Add $d_i$ as positive example to Learner

            Case 2 (YOLO Detection Validated):
                Else if $\exists d_j$ with $s_j > \tau_{conf}$:
                    $B_t \leftarrow d_j$
                    Re-initialise Tracker with $B_t$

            Case 3 (Tracker Only):
                Else if tracker confidence high:
                    $Bt \leftarrow Btrack$

            Case 4 (Failure):
                Else: $B_t \leftarrow \emptyset$

        Online Learning:
            Select negative examples from $D_{yolo}$ far from $B_t$
            Update Learner with new positive/negative examples

        $B_{prev} \leftarrow B_t$
  end for
  return $\{B_t\}$
```

4.3.2 Developed Loss Function

The key enhancement in our GAN model is a novel, composite loss function for the generator. Whilst the discriminator uses standard cross-entropy loss (Equation 4.1), the generator's loss is a weighted sum of three components and is designed to produce more potent and realistic adversarial examples.

Baseline Adversarial Loss

The foundation is the standard GAN loss, which encourages the generator to fool the discriminator. We use the non-saturating version of the log loss as follows (Goodfellow et al., 2014a):

$$\mathcal{L}_{adv} = \mathbb{E}_{z\sim p_z(z)} [\log (1 - D(G(z)))]. \tag{4.1}$$

L1 Perceptual Loss

An L1 loss is incorporated to ensure that the generated adversarial samples do not deviate excessively from the original input, thus preserving imperceptibility. This loss penalises large perturbations as follows:

$$\mathcal{L}_{L1} = E_{x\sim p_{data}(x),\, z\sim p_z(z)} [\|x - G(z)\|_1], \tag{4.2}$$

where x is a real-world sample from the data distribution p_{data}. The term $(x - G(z))$ represents the adversarial perturbation, and the L1 norm ensures sparsity and small magnitude.

Cosine Similarity Loss: The Novel Component

The primary innovation is the introduction of a cosine similarity loss. The rationale is to guide the generator to produce adversarial examples that are not just 'fooling' but also semantically similar to the real data in a high-dimensional feature space. This type of loss makes the attacks increasingly robust and transferable.

We use a pre-trained ResNet-50 (Hedeep, 2016) as a feature extractor $\phi(\cdot)$. Cosine similarity loss is defined as the cosine distance between the feature representations of the real and generated samples.

$$\mathcal{L}_{cos} = 1 - \frac{\emptyset(x).\emptyset(G(z))}{\|\emptyset(x)\|_2 \|\emptyset(G(z))\|_2} \tag{4.3}$$

Minimising $\mathcal{L}_{cos}$ ensures that the feature embeddings of G(z) and x are aligned, forcing the generator to create perturbations that alter the discriminative features for the detector (YOLO) whilst retaining the core semantic content of the original image.

Complete Generator Loss

The final loss function for the generator is a weighted sum of the three components.

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{cos}\mathcal{L}_{cos}, \tag{4.4}$$

where λ_{L1} and λ_{cos} are hyperparameters that control the trade-off amongst fooling the discriminator, perturbation imperceptibility and feature-space similarity.

4.3.3 Algorithm for Enhanced GAN Training

The training procedure for enhanced GAN is formalised in Algorithm 2.

Algorithm 4.2 Training the enhanced GAN model

```
Require: Dataset $X = \{x^{(1)}, . . . , x^{(N)}\}$, Pre-trained feature extractor $\phi$, loss weights $\lambda_{L1}$, $\lambda_{cos}$, batch size $m$
Ensure: Trained Generator $G$, Trained Discriminator $D$
1: Initialize generator $G$ and discriminator $D$ with random weights $\vartheta_G$, $\vartheta_D$
2: Initialize optimizers: $\text{opt}_G$ for $G$, $\text{opt}_D$ for $D$
3:
4: for iteration = 1 to max_iterations do
5:         // 1. Update Discriminator $D$
6:             Sample mini-batch $\{x^{(1)}, . . . , x^{(m)}\} \sim X$
7:               Sample noise vectors $\{z^{(1)}, . . . , z^{(m)}\} \sim \mathrm{N}(0, I)$
8: Generate fake samples: $x^{\tilde{}(i)} = G(z^{(i)})$ for $i = 1, . . . , m$ Line 9 - Fixed variable naming
9:           Compute discriminator loss:
10:                  $\mathcal{L}_D = -\frac{1}{m}\sum_{i=1}^{m}\left[\log D(x^{(i)}) + \log(1 - D(\tilde{x}^{(i)}))\right]$
11:             Update $D$: $\vartheta_D \leftarrow {}^{\text{opt}}{}_D(\nabla_{\vartheta D}\,\mathrm{L}_D)$
12:
13:        // 2. Update Generator $G$
```

```
14:    Sample new noise vectors $\{z^{(1)}, \ldots, z^{(m)}\} \sim \mathrm{N}(0, I)$ 15:          Generate
fake samples: $x^{\sim (i)} = G(z^{(i)})$ for $i = 1, \ldots, m$ 16:    Compute generator losses:
17:              $\mathcal{L}_{adv} = -\frac{1}{m}\sum_{i=1}^{m} \log D(\tilde{x}^{(i)})$ Non-saturating loss
18:          $\mathcal{L}_{L1} = \frac{1}{m}\sum_{i=1}^{m} \|x^{(i)} - \tilde{x}^{(i)}\|_1$
             $\mathcal{L}_{cos} = \frac{1}{m}\sum_{i=1}^{m}\left[1 - \frac{\phi(x^{(i)})\cdot\phi(\tilde{x}^{(i)})}{\| \quad ^{(i)} \|\| \quad ^{(i)} \|}\right]$
19:                                                            —      $(x\ )\ \phi(x^{\sim}\ )$ Line 15 –
Explicit cosine similarity $\phi$
20:              $\mathrm{L}_G = \mathrm{L}_{adv} + \lambda L1 \mathrm{L}_L 1 + \lambda cos \mathrm{L} cos$
21:            Update $G$: $\vartheta_G \leftarrow {}^{\mathrm{opt}}{}_G(\nabla_{\vartheta G}\, \mathrm{L}_G)$
22: end for
23: return $G$, $D$
```

4.4 Adversarial Evasion Attack Methodology

The enhanced GAN model serves as the core for generating adversarial evasion attacks. The process involves using trained generator G to perturb input frames x, creating adversarial examples x′ = G(z|x), which are then fed to the YOLO-TLD system to evaluate its robustness. This approach falls under the category of universal adversarial attacks because the generator learns a mapping function that can perturb any input.

4.5 Summary

This chapter laid the technical foundation for the thesis by presenting two key contributions.

1. The **YOLO-TLD Defense Framework**, a novel integration that enhances YOLO’s resilience to occlusions and adversarial attacks through robust tracking and online learning, formalized in Algorithm 4.1.
2. An **Enhanced GANN Attack Model**, which introduces a novel loss function combining adversarial, L1, and cosine similarity components to generate potent and semantically coherent adversarial examples, formalized in Algorithm 4.2.

The synergistic design of the two models (one for defence and one for attack) enables a comprehensive and rigorous evaluation of the proposed AI-powered surveillance system’s robustness, which will be discussed in the subsequent experimental chapter.

Chapter 5

Implementation and Testing

This chapter addresses Research Questions 2 and 3 and focuses on the implementation of the practical side of the research, namely, implementation and evaluation of the enhanced object detectors by using real and fake data. It starts with a discussion of datasets, such as COCO 2020 and VIRAT, through which important vulnerabilities in surveillance systems can be identified. Then, the chapter goes into detail about the technical framework used, including the programming language, libraries, IDE and hardware. It encompasses the application showcase developed for object tracking and the technical details to reproduce enhanced detection models, such as YOLO and SVM with HOG implementation.

The chapter continues with the application to implement the enhanced GAN model for generating fake data to simulate attacks, the evaluation criteria and results and the results showing the effect of GAN-based attacks on detection capabilities. The chapter concludes by summarising the effectiveness of GANs in launching deceptive attacks, discussing the need for further research to enhance object detectors and maintain high levels of security as the surveillance AI scales.

The research aims to apply enhanced object detectors for fake and real data (images and videos). Fake data, which emulate a malicious attack, are generated by the improved GAN model to fool the defensive system. According to the research methodology, this work is challenged by implementing it on real-time video data from surveillance cameras. The overall work is to measure the reliability of the improved detection system on the data generated by empowered adversarial attacks. These kinds of fake data are largely used in numerous areas, such as image processing, pattern and facial recognition, intrusion detection and malware detection. The chapter starts by describing the technical implementation of the work in general.

5.1 Datasets and Experimental Setup

In the expanding realm of AI, video surveillance systems leveraging AI technology are becoming increasingly prevalent. These systems offer remarkable potential for enhancing security and safety measures across myriad sectors. However, as these surveillance systems become increasingly sophisticated, so do the threats against them. This study investigated enhanced GAN-based generated attacks to understand their potential effects and explore the extent of vulnerabilities present in AI-empowered video surveillance systems.

To provide a comprehensive examination, the study utilised three distinct and robust datasets, namely, COCO 2020, VOC 2007 and VIRAT, for data collection, as shown in Figure 5.1. These diverse datasets, composed of real-world images and surveillance videos, offer an expansive range of objects, scenes and activities and form an inclusive and challenging backdrop for this study's experiments. The choice of these datasets reflects the study's intention to scrutinise the vulnerabilities of AI-empowered surveillance systems in a wide range of realistic scenarios, thereby contributing meaningful and practical insights to the field.

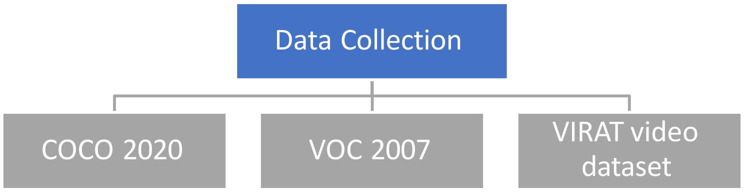


Figure 5.1 Data coordination for this research

5.1.1 COCO 2020

The newly proposed advanced detection model was applied to static images derived from the COCO 2020 dataset to substantiate the wide-ranging efficacy of the model (Lin et al., 2014). COCO is an expansive database that was primarily established to foster advancements in object detection, segmentation and captioning in the domain of computer vision.

This study employed diverse segments of the COCO dataset, including the training, validation and test sets. These segments collectively comprise more than

200,000 images that span 80 discrete object categories, ensuring a comprehensive evaluation across a multitude of instances and contexts. This complexity is further augmented by the COCO dataset's extensive features, such as object segmentation, recognition in context and superpixel stuff segmentation.

Furthermore, the COCO dataset features over 330,000 images, more than 200,000 of which are labelled. The dataset embodies approximately 1.5 million instances of objects and includes images of 250,000 individuals, each annotated with key points. This granular level of detail aids in the recognition of a broad array of object types and bolsters the model's capability to discern specific elements within large, highly complex contexts (Lin et al., 2014).

By testing the proposed model on such a comprehensive dataset, we can ascertain its effectiveness and versatility in managing a diverse range of real-world scenarios, thereby demonstrating its robust capacity for performance generalisation.

5.1.2 VOC2007

The PASCAL VOC 2007 dataset can be of substantial value as a data collection source in a study examining enhanced GAN-based generated attacks against AI-empowered video surveillance systems (Everingham et al., 2010). This dataset is a comprehensive collection of annotated images specifically designed to promote research on object detection, segmentation and recognition.

The importance of utilising the VOC 2007 dataset for such a study is manifold. Firstly, the dataset offers a broad array of image categories that cover various aspects of everyday scenes. This diversity is crucial when creating and testing an AI model capable of recognising a wide range of objects and scenarios, ensuring that the model is adequately trained to understand and categorise different types of behaviours and objects in surveillance videos. Secondly, the VOC 2007 dataset provides detailed annotations. These annotations include object classes and specific bounding box coordinates for each object in an image. This annotated information is beneficial for the supervised learning of GAN models. Such a supervised approach helps the GAN model in learning to generate accurate and complex attacks, making the assessment highly comprehensive and informative (Everingham et al., 2010). Finally, the dataset's standardised nature allows for a straightforward comparison of the developed model’s performance against that of other models trained on the same data. This comparison contributes to the validity and transparency of the study, enhancing its overall influence.

In summary, the VOC 2007 dataset, with its diverse set of annotated images and standardised data structure, is an invaluable resource for this study because it facilitates a comprehensive evaluation and refinement of the GAN-based model. The images include at least one object of the corresponding class, and the number of object instances is presented in Table 5.1. The data are divided into training (train)/validation (val) and testing (test) datasets.

Table 5.1 Data coordination for this research

Object type	**Train**		**Val**		**Train-val**		**Test**	
	Images	**Objects**	**Images**	**Objects**	**Images**	**Objects**	**Images**	**Objects**
Aeroplane	112	151	126	155	238	306	204	285
Bicycle	116	176	127	177	243	353	239	337
Bird	180	243	150	243	330	486	282	459
Boat	81	140	100	150	181	290	172	263
Bottle	139	253	105	252	244	505	212	469
Bus	97	115	89	114	186	229	174	213
Car	376	625	337	625	713	1,250	721	1,201
Cat	163	186	174	190	337	376	322	358
Chair	224	400	221	398	445	798	417	756
Cow	69	136	72	123	141	259	127	244
Dining Table	97	103	103	112	200	215	190	206
Dog	203	253	218	257	421	510	418	489
Horse	139	182	148	180	287	362	274	348
Motorbike	120	167	125	172	245	339	222	325
Person	1,025	2,358	983	2,332	2,008	4,690	2,007	4,528
Potted plant	133	248	112	266	245	514	224	480
Sheep	48	130	48	127	96	257	97	242
Sofa	111	124	118	124	229	248	223	239
Train	127	145	134	152	261	297	259	282
TV/monitor	127	166	128	156	256	324	229	208
Total	**2,501**	**6,301**	**2,510**	**6,307**	**5,011**	**12,608**	**4,952**	**12,032**

5.1.3 VIRAT Video Dataset

The VIRAT video dataset is a crucial source of data for a study focusing on enhanced GAN-based generated attacks against AI-empowered video surveillance systems (Oh et al., 2011). VIRAT is a groundbreaking, large-scale dataset that has been extensively used for research on activity recognition and anomaly detection in video data. It is a publicly available database that encompasses video sequences collected in realistic, uncontrolled outdoor environments. It includes a wide variety of activities and events, along with annotations that identify the activities and objects present. VIRAT, developed by the US National Institute of Standards and Technology, plays an integral role in advancing the field of video surveillance.

The importance of using the VIRAT dataset in this context is manifold. Firstly, VIRAT provides a wide spectrum of video sequences that have been recorded in various real-world environments, making it an ideal platform for developing and evaluating methods designed to understand, identify and categorise behaviours and events occurring in video data (Oh et al., 2011).

The dataset includes various features, such as object detection, tracking, event recognition, activity understanding and anomaly detection. Specifically, it contains annotations for object detection and tracking (including people and vehicles) and event annotation at different levels of the activity hierarchy. Events are classified into single-object, two-object interactive and multi-object events, providing a wide range of scenarios for researchers to test their models.

Moreover, the dataset presents complex scenarios with multiple events and interactions, allowing the model to be tested against highly complicated, intricate attacks. Given the focus on AI-empowered video surveillance systems, the variety and complexity of the data in the VIRAT dataset offer a robust platform for testing the resilience of these systems under different adversarial scenarios.

VIRAT also comprises videos captured in realistic outdoor settings (e.g. parking lots and shopping malls). It encompasses a range of activities and events observed in these environments. The videos were recorded from different angles and viewpoints, providing researchers with various challenging scenarios for testing their models.

Moreover, with its annotated labels and identified events, the VIRAT dataset enables supervised training for GAN models. It allows an enhanced GAN-based model to learn and generate precise attacks, thereby improving the study's results and conclusions.

In summary, the VIRAT dataset's breadth, complexity and detailed annotation make it a vital source of data for this study. It provides the necessary conditions for rigorous evaluation and enhancement of GAN-based models because VIRAT dataset's realistic, diverse and richly annotated nature is instrumental in training computer vision models to handle complex, real-world situations, leading to advancements in activity recognition and anomaly detection.

5.1.4 Comparison of Data Collection Methods

This part delves into the heart of our methodology by discussing in detail the datasets utilised for our investigation: COCO 2020, VOC 2007 and VIRAT. The three datasets were instrumental in our exploration of enhanced GAN-based generated attacks against AI-empowered video surveillance systems. Each dataset, with its unique composition and variety of instances, offered distinct advantages that collectively enriched the breadth and depth of our research. This section provides a detailed comparison of these datasets by discussing their attributes and explaining how their individual characteristics contributed to the thoroughness and robustness of our study. This comparison underscores the distinct advantages of each dataset and elucidates why the combination of all three was crucial for our comprehensive understanding of the vulnerabilities of AI-empowered video surveillance systems. Table 5.2 shows a comparison of the three datasets, bringing to light the unique and collective strengths they brought to our research.

In summary, the comprehensive use of COCO 2020, VOC 2007 and VIRAT video datasets substantially contributed to our understanding of enhanced GAN-based generated attacks against AI-empowered video surveillance systems. These datasets, each featuring a distinct array of images and videos, provided an expansive and realistic context for our investigation, thereby enhancing the relevance and applicability of our findings.

Table 5.2 Comparison of the methods used for data collection

	COCO 2020	**VOC 2007**	**VIRAT Video Dataset**
Data Type/Format	Images	Images	Videos

Primary Purpose	Image recognition, object detection, semantic segmentation, captioning	Image classification, object detection, semantic segmentation	Activity recognition, anomaly detection
Annotations	Objects (~1.5 million), segmentation masks, key points, captions	Objects (~27 k), segmentation masks (27,450 ROIs in 9,963 images)	Activities, objects, people (12+ hours of video, numerous instances)
Size	~200,000 images	~20,000 images	~8.5 hours of video
Year Released	2020	2007	2011

By utilising COCO 2020 and VOC 2007, we explored a wide range of objects and scenes, which added depth to our analysis of attacks in various scenarios. The VIRAT video dataset proved invaluable in simulating real-world surveillance situations, playing a critical role in assessing the robustness of AI-empowered systems against GAN-based attacks.

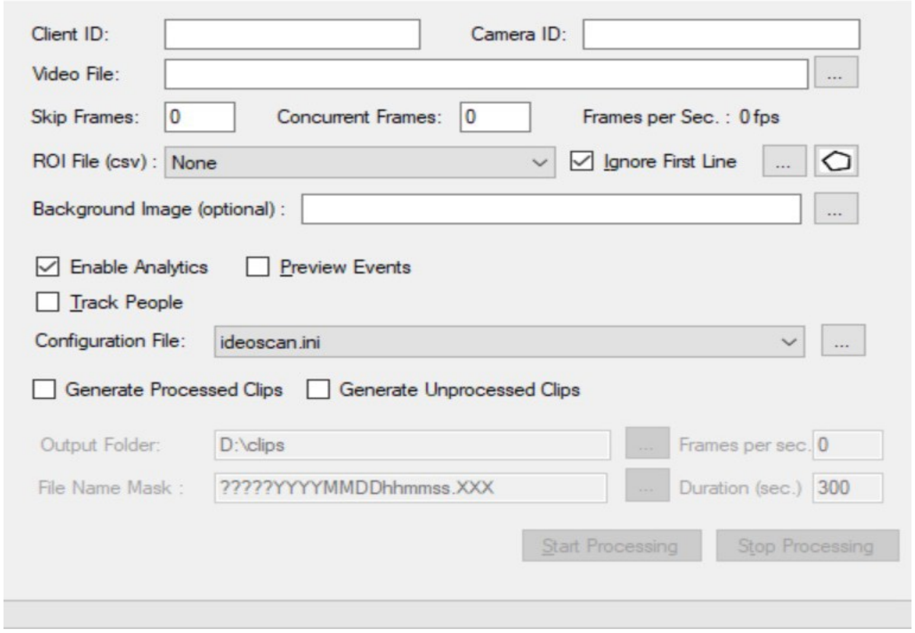
Through our study, we demonstrated the potential threats and vulnerabilities present in AI-powered video surveillance systems. However, by the same token, we have also paved the way for future research to fortify these systems and further harness the capabilities of AI in surveillance. The datasets used in this research have proven to be robust platforms for assessing these threats. We hope that future work will continue to leverage such resources to enhance the security and effectiveness of AI-empowered video surveillance systems.

Although our research provides substantial insights, it is but one step in the continuous journey of learning and improvement in the realm of AI security. We anticipate that the techniques and insights derived from this study will serve as a foundation for future investigations, leading to other robust and secure AI-empowered video surveillance systems that can withstand even the most sophisticated attacks.

5.2 Implementation Framework

The chosen framework for the application’s front end, which will demonstrate the work accomplished in this project, is C#. The reason is that it requires far fewer configurations than other frameworks and does not concern itself with the project having various apps, making it easy and straightforward to use.

Figure 5.2 presents the developed tool for object tracking with a video source. Currently, the tool runs on recorded videos to demonstrate the algorithm's ability and accuracy rather than being connected to real-life videos. Multiple cameras can be connected because each surveillance camera has its own ID. The output file is saved in local storage for analysis and verification of the applied algorithm.



Client ID: Camera ID:

Video File:

Skip Frames: 0 Concurrent Frames: 0 Frames per Sec. : 0 fps

ROI File (csv) : None ☑ Ignore First Line

Background Image (optional) :

☑ Enable Analytics ☐ Preview Events

☐ Track People

Configuration File: ideoscan.ini

☐ Generate Processed Clips ☐ Generate Unprocessed Clips

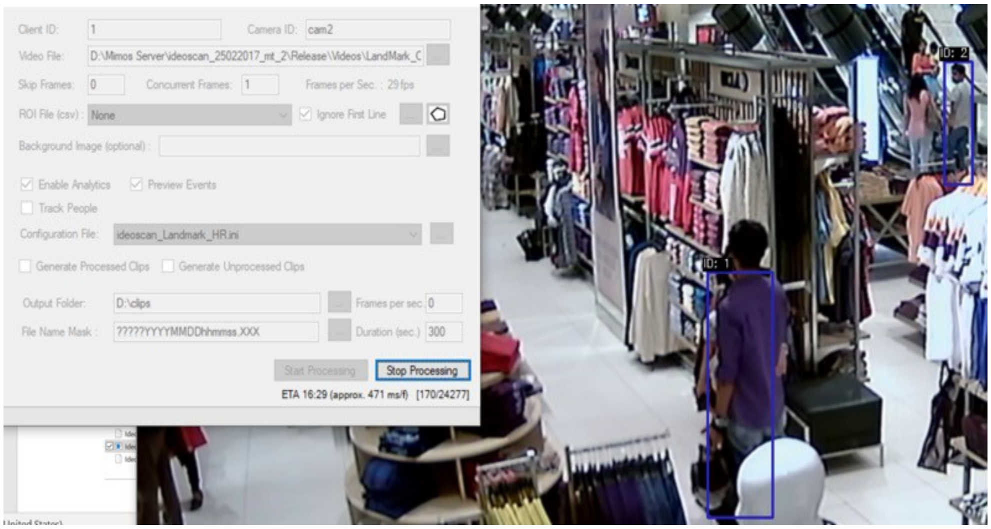
Output Folder: D:\clips Frames per sec. 0

File Name Mask : ?????YYYYMMDDhhmmss.XXX Duration (sec.) 300

Figure 5.2 Screenshots of the developed front-end application

Figure 5.3 shows that two people are detected by the surveillance camera ID Cam2. The two people are given IDs after about 0.4 milliseconds of processing time (this time depends on the used hardware).

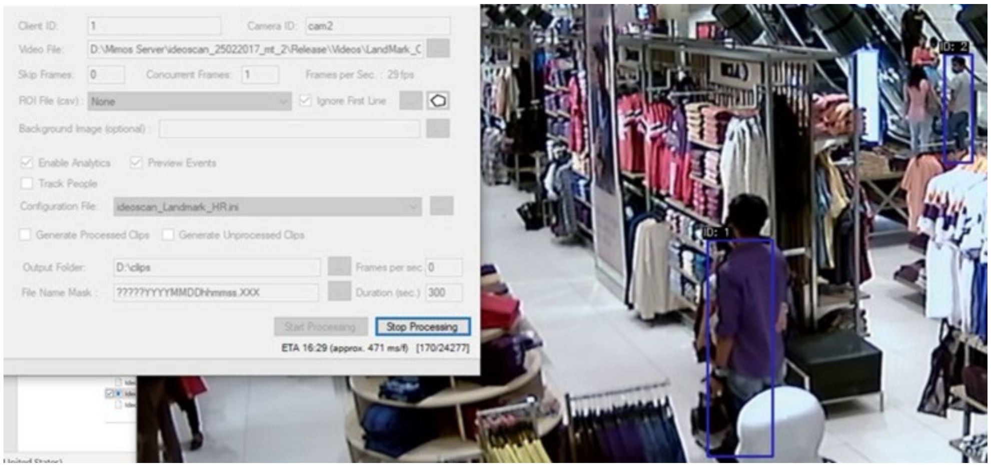
The given IDs are assigned to the two people even when they disappear from the screen but under two conditions: they are still in a nearby location or the maximum time lapse is 4 seconds. The developed algorithm works with normal human behaviour that has malicious activity. This step benefits GAN analysis because it keeps track of the same person or object that appeared in the camera.



Client ID: 1
Camera ID: cam2
Video File: D:\Mimos Server\ideoscan_25022017_mt_2\Release\Videos\LandMark_C
Skip Frames: 0
Concurrent Frames: 1
Frames per Sec. : 29 fps
ROI File (csv) : None
☑ Ignore First Line
Background Image (optional) :
☑ Enable Analytics ☑ Preview Events
□ Track People
Configuration File: ideoscan_Landmark_HR.ini
□ Generate Processed Clips □ Generate Unprocessed Clips
Output Folder: D:\clips
Frames per sec. 0
File Name Mask : ?????YYYYMMDDhhmmss.XXX
Duration (sec.) 300
ETA 16:29 (approx. 471 ms/f) [170/24277]

Figure 5.3 TLD output of running the system after 0.4 milliseconds

Figure 5.4 shows that Person 1 is walking, and the given ID (ID 1) remains the same. However, the second person behind Person 1 is not detected yet but will be detected after having a clear movement and an image for the person; a new ID will be assigned as well.



Client ID: 1
Camera ID: cam2
Video File: D:\Mimos Server\ideoscan_25022017_mt_2\Release\Videos\LandMark_C
Skip Frames: 0
Concurrent Frames: 1
Frames per Sec. : 29 fps
ROI File (csv) : None
☑ Ignore First Line
Background Image (optional) :
☑ Enable Analytics ☑ Preview Events
□ Track People
Configuration File: ideoscan_Landmark_HR.ini
□ Generate Processed Clips □ Generate Unprocessed Clips
Output Folder: D:\clips
Frames per sec. 0
File Name Mask : ?????YYYYMMDDhhmmss.XXX
Duration (sec.) 300
ETA 16:29 (approx. 471 ms/f) [170/24277]

Figure 5.4 TLD output of running the system after 0.6 milliseconds

5.3 Model Implementation and Testing

This work is expected to face challenges in implementing AI algorithms for object detection on static images and real-time video data from surveillance cameras. For this reason, the implementation was divided into two main stages, aligning with the one proposed in the methodology.

- Selection and enhancement of the detection model
- Enhanced GAN model

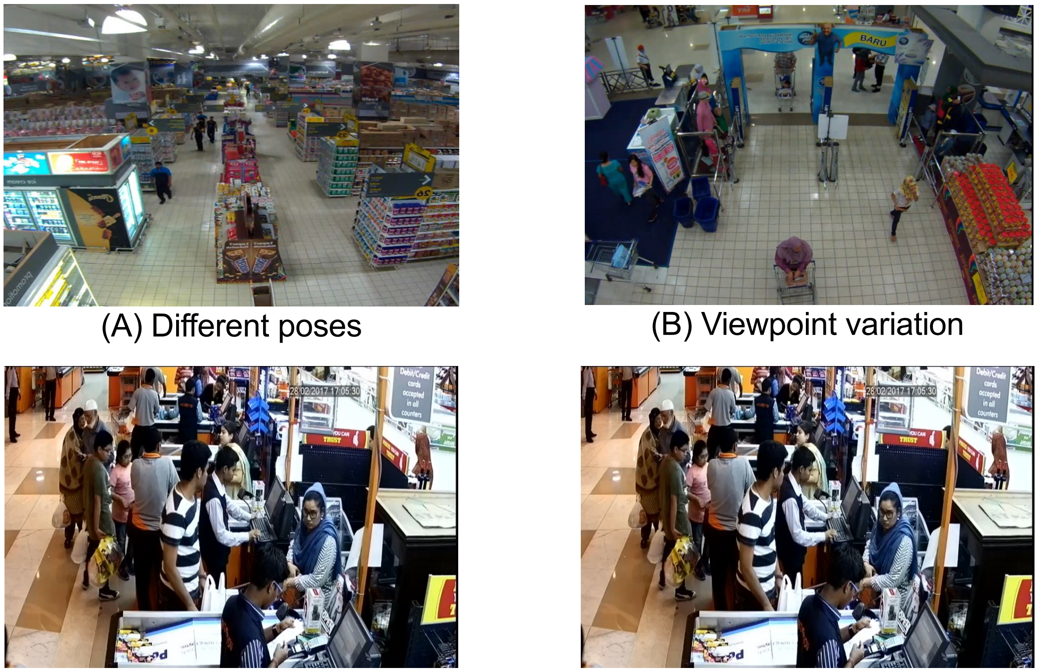
5.3.1 Detection Model Selection and Enhancement

The work began by evaluating various detection models on different objects. The implementation of different people detection algorithms was conducted first. The implemented algorithms were Haar cascade, YOLO, KCF and SVM with HOG features. HAAR cascade is an object detection algorithm used to identify faces in an image or a real-time video. The algorithm uses edge or line detection features proposed by Viola and Jones in their paper 'Rapid Object Detection Using a Boosted Cascade of Simple Features'.

In comparison with Haar cascade, YOLO is a single-step approach for object detection and classification; the class is predicted after the bounding box is evaluated for the input. Several variations of YOLO have been explored. Recently, YOLOv5 was improved to detect small objects efficiently and precisely by integrating DeepSORT for multi-objects. YOLOv5 is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images.

YOLO uses features learned by a deep CNN to detect an object. Versions 1–3 of YOLO were created by Joseph Redmon and Ali Farhadi (Sreeja et al., 2023). YOLO is a CNN for performing object detection in real time. CNNs are classifier-based systems that can process input images as structured data arrays and identify patterns between them. YOLO performs much faster than other networks whilst maintaining accuracy. KCF has also been tested, with recent studies showing promising results. Based on the idea of a traditional correlational filter, KCF utilises the kernel trick and circulant matrices to remarkably improve computation speed. The main advantage of KCF is that it can be trained independently by using the first consecutive frame, which reduces computational complexity for object detection and tracking. Meanwhile, SVM, a type of ML, is used to identify people on the basis of HOG. HOG features mainly describe an object's structure or shape, in addition to the edge orientation, on the basis of the magnitude and direction of extracted information.

All the techniques mentioned above were evaluated in terms of accuracy and timing to determine the most effective amongst them. The chosen method was used for the next step, which involved object detection for objects carried by persons. One hundred frames featuring approximately 332 people were used for the evaluation stage. These frames were collected from videos that present various challenges, including different viewpoint variations, pose variations, occlusions and objects (e.g. mannequins).



(A) Different poses

(B) Viewpoint variation

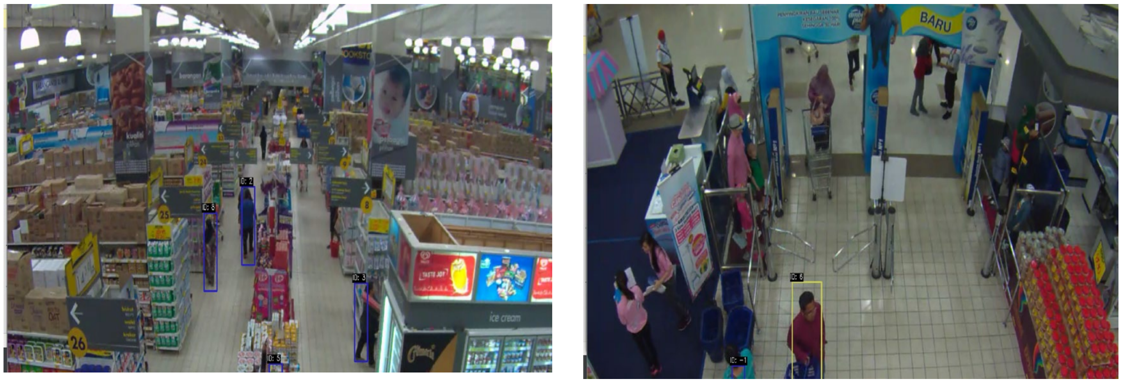
(C) Occlusion

(D) Objects like humans

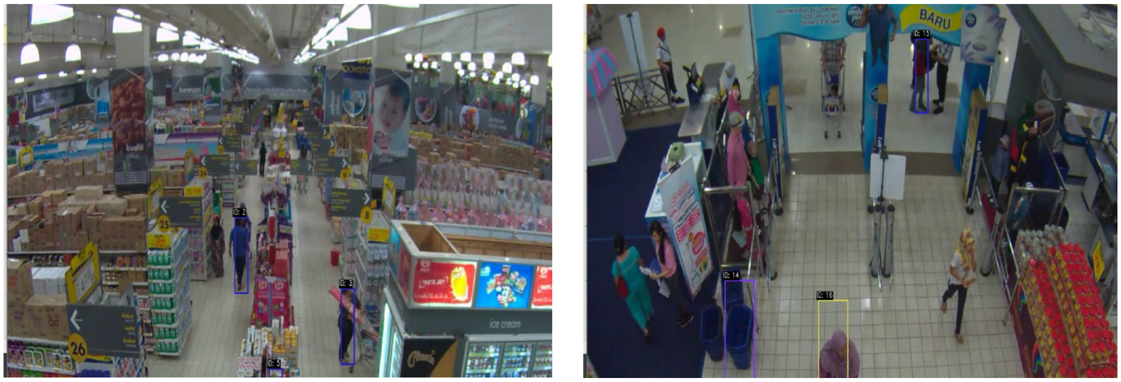
Figure 5.5 Samples of images captured from videos with different challenges

Table 5.3 Performance evaluation (time and accuracy) of different detection models

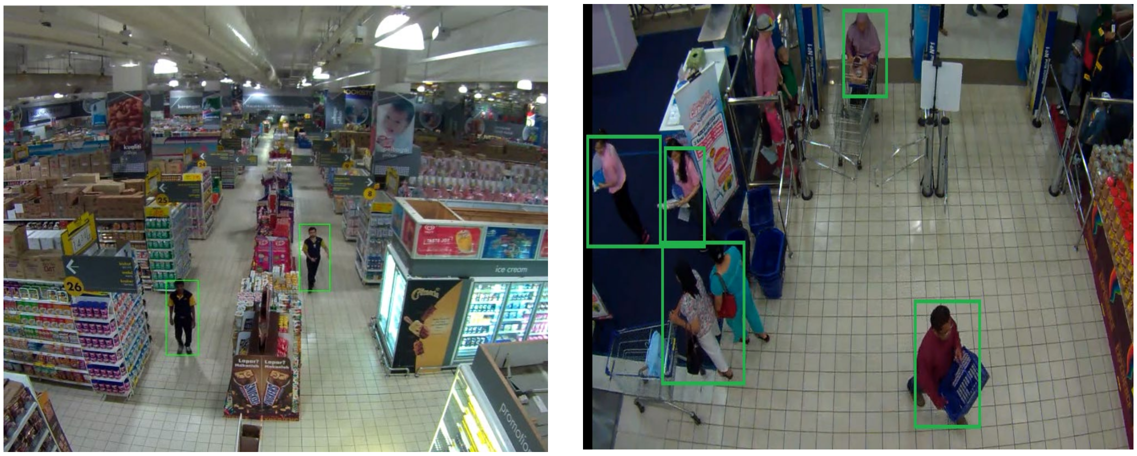
	HAAR	KCF	SVM	YOLO
Accuracy	30.12% (100/332)	73.7% (245/332)	78.9% (262/332)	92.7% (308/332)
Timing per frame	332 ms	214 ms	282 ms	387 ms



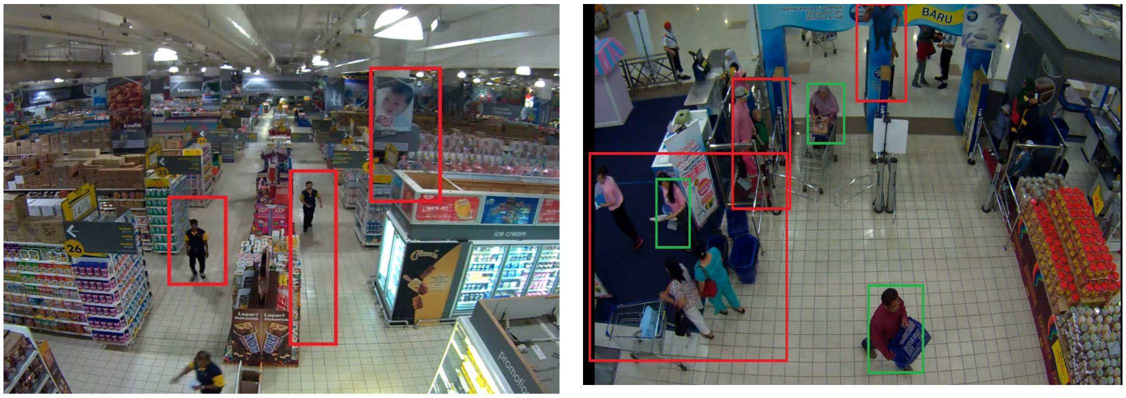
(a) KCF detection samples



(b) Haar Cascade detecion sample



(c) Yolo detection sample



(d) SVM with HOG detection sample

Figure 5.6 Visual performance of different models (a sample image snapped from a video)

As shown in Table 5.3 and Figure 5.6, YOLO outperforms the other techniques that are based on ML and image processing. The possibility of enhancing the performance of YOLO whilst processing videos with multiple frames by adding a reliable tracker has been explored. People tracking is achieved by detecting and tracking individuals within a defined ROI or zone.

ROI needs to be defined at the first stage, followed by the detection of a person within this ROI and assigning them a unique ID, which is then counted and tracked. A new entry in this ROI is addressed in a similar manner. In the event of the exit of one of the persons or entries, that person is no longer tracked or counted by ID. TLD is an award-winning, real-time algorithm for tracking objects in video streams. The object of interest is defined by a bounding box in a single frame. TLD simultaneously tracks the object, learns its appearance and detects it whenever it appears in the video. The result is real-time tracking that typically improves over time.

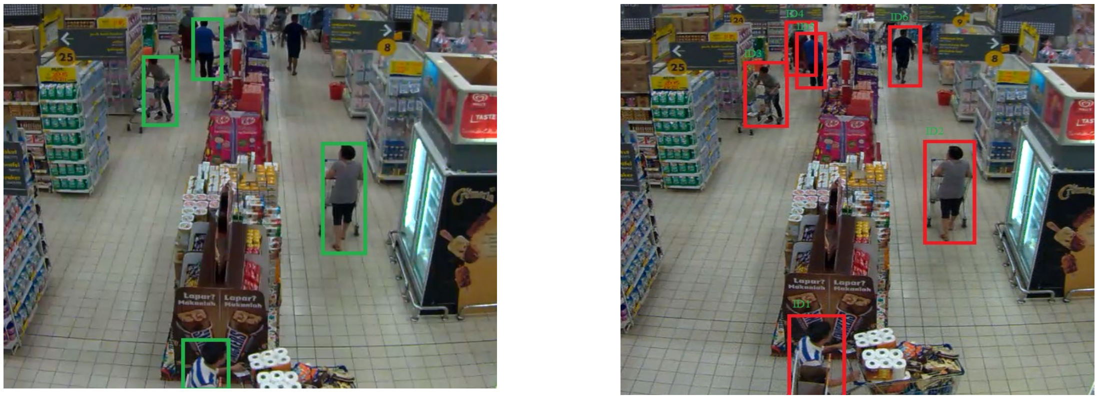
The main features of TLD are as follows:

– TLD can currently track multiple objects in parallel;
– No offline training stage;
– Real-time performance.

The implication of adding TLD to help YOLO perform well in detecting people is clarified in the table below.

Table 5.4 Improved YOLO accuracy over a sample video

	YOLO	YOLO with TLD
Accuracy	92.7% (308/332)	96.38 (320/332)
Timing per frame	387 ms	402 ms



(1) 4/6 people detected YOLO only

(2) 6/6 people detected by YOLO + TLD

Figure 5.7 Basic vs. improved YOLO performance (a sample image snapped from a video)

As shown in Table 5.4 and Figure 5.7, TLD-tracker enhanced YOLO produces better results than YOLO with no TLD tracking in detecting targeted objects without incurring a major surge in processing power. This finding highlights the effectiveness of the new enhancement of YOLO in efficiently detecting images.

Figure 5.8 COCO dataset consisting of 123,287 images

The proposed enhanced detection model was also applied to static images from the COCO 2020 dataset to confirm the generalisation of its performance. COCO is a large-scale object detection, segmentation and captioning dataset. The COCO training, validation and test sets contain more than 200,000 images and 80 object categories with several features (e.g. object segmentation, recognition in context and superpixel stuff segmentation). Specifically, it covers 330 K images (>200 K labelled), 1.5 million object instances, 80 object categories and 250,000 people with key points. The enhanced model calculated and evaluated the detection of various objects (Figure 5.9), including mobile phones, backpacks, motorbikes and weapons as shown in Table 5.5.

Table 5.5 Improved YOLO accuracy detection over objects from the COCO dataset

	Mobile phones	Vehicles	Motorbikes	Weapons
YOLO with TLD	68.4%	8.88%	82.65%	61.34

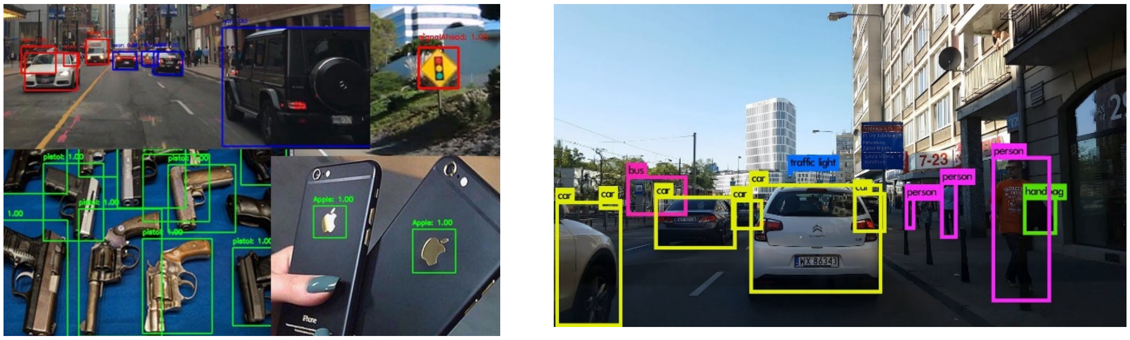


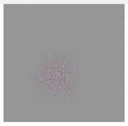
Figure 5.9 Improved YOLO visual performance over different objects by using the COCO20 dataset

5.3.2 Enhanced GAN Model

The main objective of the conducted research is to apply enhanced object detectors to fake and real data (images and videos). Fake data, which emulate a malicious attack, are generated by the improved GAN model to fool the defensive system. This work is challenged by implementing it on real-time video data from surveillance cameras in accordance with the research methodology.

The overall aim of the work is to measure the reliability of the improved detection system on data generated by empowered adversarial attacks. These fake data are largely used in numerous areas, such as image processing, pattern and facial recognition, intrusion detection and malware detection. The evaluation results are discussed in the next chapter. Table 5.6 presents the initial preliminary results of the GAN attacks on some images.

Table 5.6 Sample images snipped from the video detected and recognised by the enhanced YOLO algorithm

Samples	Clean Images	Added Noise	Classifications		
			Before Attack	After Attack	Detection Accuracy
1			Person	Bag	0.67
2			Person	Bag	0.71
3			Car	ND*	1

*ND means ‘not detected’.

GANs are neural network architectures used to generate new data samples that are similar to a training dataset. GANs consist of two models: a generator and a discriminator.

The generator produces new samples, and the discriminator tries to distinguish the generated samples from the real ones. During training, the generator and discriminator are trained simultaneously, with the generator attempting to produce samples that are difficult for the discriminator to distinguish from real ones, and the discriminator attempting to correctly classify the generated samples as fake.

GAN attack on images implemented using Keras and TensorFlow 2 requires defining and compiling the generator and discriminator models and training them by using a dataset of images. A general outline of the steps to follow is given below.

- Import the necessary libraries, such as *tensorflow*, *keras* and *numpy*.
- Define the generator model. This model should adopt a random noise vector as an input and produce an image. A combination of convolutional and transposed convolutional layers can be used to up-sample the noise vector and generate an image.

- Define the discriminator model. This model should use an image as an input and produce a single scalar value representing the probability that the image is real. A combination of convolutional and dense layers can be employed to extract features from the image and make a classification decision.
- Compile the generator and discriminator models by using an appropriate loss function and optimiser.
- Define a function to train GAN. This function should iterate over several epochs, and at each epoch it should:
  a. Generate a batch of random noise vectors.
  b. Use the generator to produce a batch of fake images from the noise vectors.
  c. Combine the fake images with a batch of real images from the training dataset.
  d. Train the discriminator on this combined batch, with the fake images labelled as 'fake' and the real images labelled as 'real'.
  e. Generate a new batch of random noise vectors.
  f. Use the generator to produce a batch of fake images from the noise vectors.
  g. Train the generator by using the discriminator's predictions on the fake images as the label.
- Load a dataset of images and use it to train GAN via the function defined in Step 5.

However, additional processes, such as the specific architecture of the generator and discriminator, the size and shape of the input and output tensors and the hyperparameters of the models, are considered. Furthermore, dataset pre-processing, such as normalising the pixel values or resizing the images to a consistent size, is performed before training.

The PASCAL VOC dataset is a widely used dataset for image classification and object detection tasks. It consists of images from various real-world scenarios and annotated with bounding boxes and class labels for various objects, such as aeroplanes, cars and pedestrians. The PASCAL VOC 2007 dataset consists of 5,011 images, which are divided into a training set and a validation set and organised into 20 classes. The dataset is commonly used to train and evaluate object detection models.

The training set contains 4,952 images, and the validation set contains the remaining 59 images. The images in the dataset are annotated with bounding boxes and class labels for various objects, such as aeroplanes, cars and pedestrians. The objects belong to 20 classes, namely, aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train and TV/monitor.

5.4 Summary

Chapter 5 addressed Research Question 2 by demonstrating the implementation and evaluation of the enhanced GAN model for generating adversarial samples that simulate malicious attacks on surveillance systems. The chapter assessed the effect of these GAN-based attacks on detection capabilities by using datasets, such as COCO and VIRAT. By analysing the results of GAN attacks on enhanced object detectors (YOLO and SVM with HOG), it highlighted the extent to which fake data can compromise these systems.

The chapter also answered Research Question 3 by providing an empirical evaluation of the vulnerabilities of surveillance systems through the application of adversarial GAN-based attacks. It examined the effectiveness of existing detection systems and highlighted the necessity for robust defensive mechanisms. By implementing enhanced object detectors and evaluating their performance under adversarial conditions, the chapter shed light on current system weaknesses and the need for continued research in this area.

The implementation aspects of the detection system mode, as well as the enhanced GAN model, within the suggested framework were summarised. The chapter included dataset selection and usage, the technical setup of the framework domain and the application and testing of the proposed enhanced GAN model. For data collection, useful and popular image datasets, including COCO, VOC2007 and VIRAT, were selected. Their integration into the training and testing of the developed models was analysed. The rationale of their benefits for evaluating GAN attacks on the detection systems was identified.

The technical framework part described the programming languages, libraries and development environment that facilitated the easy integration of models and data. It also included the hardware configuration, which was crucial for executing the computationally intense GAN-based attacks and the YOLO-based detection systems.

In general, this chapter demonstrated the technical feasibility of using adversarial attack methods and GAN models to compromise video surveillance systems, highlighting the necessity for developing reliable defensive mechanisms.

Chapter 6

Performance Evaluation

This chapter presents a comprehensive performance evaluation of the proposed GAN-based adversarial attack model and the enhanced YOLO-TLD defence system, addressing Research Questions 2 and 3. The objective is to quantitatively assess the robustness of the AI-empowered video surveillance framework under adversarial conditions. Through standardised metrics, the framework's detection accuracy and processing efficiency are evaluated through controlled experiments on diverse datasets.

These experiments are designed to measure the effectiveness of the GAN model in generating successful deceptive samples and the resilience of the enhanced YOLO classifier in maintaining accurate object detection against adversarial perturbations. The results provide empirical validation of the proposed methods, offering insights into the defensive capabilities of the YOLO4-based surveillance system in real-world contexts and demonstrating the high efficiency of the enhanced GAN model in compromising the system.

6.1 Evaluation Criteria and Results

Addressing Objectives 2 and 3, we emulated malicious attacks by using GAN-generated fake objects and evaluated their effect on the detection capability of TLD-enhanced YOLO. Malicious attacks were emulated using an optimised GAN model, which generated fake objects and adversarial video frames designed to deceive the YOLO classifier. These adversarial samples were integrated into real-world datasets, such as COCO 2020 and VIRAT, simulating attacks on AI-powered surveillance systems. The effect of these attacks was evaluated by measuring detection accuracy, false positive rates and processing efficiency, revealing the vulnerabilities of the YOLO system. For enhanced defence, the YOLO classifier was fortified with a TLD module, improving its ability to detect adversarial samples. Controlled experiments validated the effectiveness of the fortified YOLO system, with metrics, such as precision and recall, demonstrating substantial improvements in resilience against adversarial attacks. This systematic evaluation highlighted the need for robust defensive mechanisms to counter GAN-generated threats effectively.

Table 6.1 summarises the outcomes of the detection before and after the GAN-based attack for images. This attack consisted of feeding the GAN model with sample images that generated similar fake images. Then, these images were subjected to a YOLO detector, and detection accuracy was measured.

Column 6 shows the average detection accuracy before the GAN-generated attack, and Column 7 exhibits the average adversarial detection accuracy after the GAN attack. The objects used were obtained from the VOC2007 dataset. The decrease in the detection accuracy of YOLO was obvious and is given in Column 8 of the table as an accuracy drop percentage. This decrease indicates serious detection failure by the enhanced YOLO, which peaked at about 95% detection failure for birds.

Table 6.1 GAN performance verification for images using the VOC2007 dataset

No.	Real Images	No. of Test Images	No. of Test Objects	Classifications (YOLO)			
				Objects Detected	Average Detection Accuracy	Average Adversarial Detection Accuracy (Fake Image)	Accuracy Drop
1	Person	2,007	4,528	Person	97.65%	11.71%	85.94%
2	Car	721	1,201	Car	95.58%	8.60%	86.98%
3	Motorbike	222	325	Motorbike	95.82%	3.34%	92.48%
4	Dog	418	489	Dog	93.15%	3.51%	89.64%
5	Airplane	204	285	Airplane	96.08%	14.66%	81.42%
6	Boat	172	263	Boat	82.84%	7.82%	75.02%
7	Cat	322	358	Cat	94.54%	2.51%	92.03%
8	Sofa	223	239	Sofa	90.96%	4.09%	86.87%
9	Bus	174	213	Bus	96.91%	8.11%	88.80%
10	Bird	283	459	Bird	95.66%	1.52%	94.14%
Average Accuracy					**93.92%**	**6.59%**	**87.33%**

Table 6.2 shows the efficiency of the GAN attack on the same dataset (VOC2007) by presenting its fooling rate (using Equation (3.1)) and the resulting efficiency drop (using Equation (3.2)) per object category. This table indicates the extent to which the GAN attack caused tremendous harm by generating a large number of fake images that evaded detection by the TLD-enhanced YOLO classifier.

Table 6.2 Fooling rate efficiency of GAN attack on the VOC2007 dataset

No.	Real Images	No. of Test Images	No. of Test Objects	Classifications (YOLO)							
				Original Images				GAN Images			Attack rate
				Detected Object Type	Detected Objects	Correctly Detected	Efficiency of Success Rate	Adversarial Detected Objects (Fake Image)	Correct Object	Efficiency of Fooling Rate	Efficiency Drop
1	Person	2,007	4,528	Person	5,220	4,287	94.68%	3,581	156	96.55%	96.36%
2	Car	721	1,201	Car	1,333	1,091	90.84%	263	77	93.59%	92.94%
3	Motorbike	222	325	Motorbike	315	315	96.92%	84	39	88.00%	87.62%
4	Dog	418	489	Dog	460	448	91.62%	126	51	89.57%	88.62%
5	Airplane	204	285	Airplane	266	264	92.63%	82	32	88.77%	87.88%
6	Boat	172	263	Boat	267	260	98.86%	88	19	92.78%	92.69%
7	Cat	322	358	Cat	321	317	88.55%	102	26	92.74%	91.80%
8	Sofa	223	239	Sofa	245	237	99.16%	66	34	85.77%	85.65%
9	Bus	174	213	Bus	219	210	98.59%	75	11	94.84%	94.76%
10	Bird	283	459	Bird	412	412	89.76%	40	8	98.26%	98.06%
								Average efficiency		**92.09%**	**91.64%**

The VIRAT video dataset was chosen to establish a standard for comparison for GAN attacks on video streaming. This dataset consists of a set of surveillance videos that accurately depict real-life CCTV recordings. The videos feature stationary cameras that capture public areas, such as sidewalks, streets and parking lots. The video resolutions range from high definition (720p) to full high definition (1080p). The videos contain various objects, but the primary ones are cars and humans. To prove the real-time capacity of the proposed system on videos captured by CCTV and to measure performance, we calculated the inference time for each video frame in milliseconds and converted it into an FPS value for benchmarking, as shown in Table 6.3.

Table 6.3 Time of object detection on the VIRAT video dataset

VIDEO ID	NO. OF FRAMES	PROCESSING TIME (SECONDS)	AVERAGE FPS
VIRAT_S_ 000002	9,074	242	37.68
VIRAT S 050000 05 000696 000732	1,070	40	27.06
VIRAT S 010200 03 000470 000567	2,090	50	42.06
VIRAT S 050203 07 001288 001531	7,280	253	28.93
VIRAT S 040104 02 000459 000721	7,841	208	38

The experiment was conducted using the hardware mentioned in the hardware section. Table 6.4 shows the different GAN settings' efficiency in correct object detection and the fooling rate for the VIRAT Video dataset. Addressing Objectives 3 and 4, we emulated malicious attacks on the videos by using GAN-generated fake objects and evaluated their effect on the detection capability of TLD-enhanced YOLO. Different GAN settings were implemented to show the influence of the generated images on the proposed TLD-enhanced detector. The last four columns of Table 6.4 present the details of the accuracy and how much is affected by the fake images generated by GAN. The fooling rate mAP is 63% for the 25 test video samples.

Table 6.4 shows a quantitative evaluation of the proposed GAN model's effectiveness in deceiving an object detector across various video sequences from the VIRAT dataset. The results demonstrated the vulnerability of a standard object detection system to adversarial attacks. The baseline detection accuracy on clean, unaltered frames was consistently high, exceeding 92% across all tested videos. However, when subjected to adversarial perturbations generated by the GAN model under four different configurations (S1–S4), the detection rate plummeted dramatically. For instance, in the first video sequence (VIRAT_S_000002), the detection rate decreased from 94.6% to as low as 3.0% under the most effective attack setting (S4). This pattern was consistent across all scenarios, confirming that the GAN model successfully generated imperceptible perturbations that substantially degraded the performance of the detection system, with the attack's intensity being controllable through specific model parameterisations (S1–S4). The table effectively benchmarks the attack strength of different GAN configurations, providing critical insights for developing future defence mechanisms.

Table 6.4 Detection and fooling rate for the VIRAT dataset under different GAN settings

Video ID	No. of Frames	No. of Persons	No. of Vehicles	Detection Rate (Real Frame)	Detection Rate GAN (S1)	Detection Rate GAN (S2)	Detection Rate GAN (S3)	Detection Rate GAN (S4)
VIRAT_S_000002	9,074	10,255	40,833	94.6%	19.4%	16.2%	10.3%	3.0%
VIRAT_S_050000_05_000696_000732	1,070	242	4,322	95.3%	19.0%	15.9%	10.5%	7.4%
VIRAT_S_010200_03_000470_000567	2,090	824	3,962	94.8%	21.3%	17.7%	10.8%	5.5%
VIRAT_S_050203_07_001288_001531	7,280	3,625	10,690	92.6%	12.8%	11.1%	8.1%	3.8%
VIRAT_S_040104_02_000459_000721	7,841	4,269	9,525	94.0%	14.7%	12.9%	7.4%	3.0%

*Note: S1-S4 represent four distinct parameter configurations for the GAN model.

6.2 Evaluation of the GAN Model

The results were compared in terms of the achieved loss function MSE for the generator and discriminator during training. The training termination factors were set to either reaching the maximum number of epochs (100) or achieving the desired performance with no change (more or less than 0.1). Figure 6.1 shows that the generator obtained the ability to produce adversarial images at the global minimum state. The steadily decreasing generator loss showed that the generator was learning to create outputs that were increasingly effective in fooling the discriminator. The loss function also performed very well because the learning curve is not extremely low or near zero, which may be a sign of mode collapse, where the generator produces limited output. Mode collapse has been explained by Li et al. (2021a).

Furthermore, Figure 6.2 illustrates the discriminator's ability to distinguish between authentic and generator-produced adversarial data. A moderately fluctuating discriminator loss curve is ideal because it indicates an adaptive and dynamic training process where the generator continuously pushes the discriminator to enhance its detection capabilities.

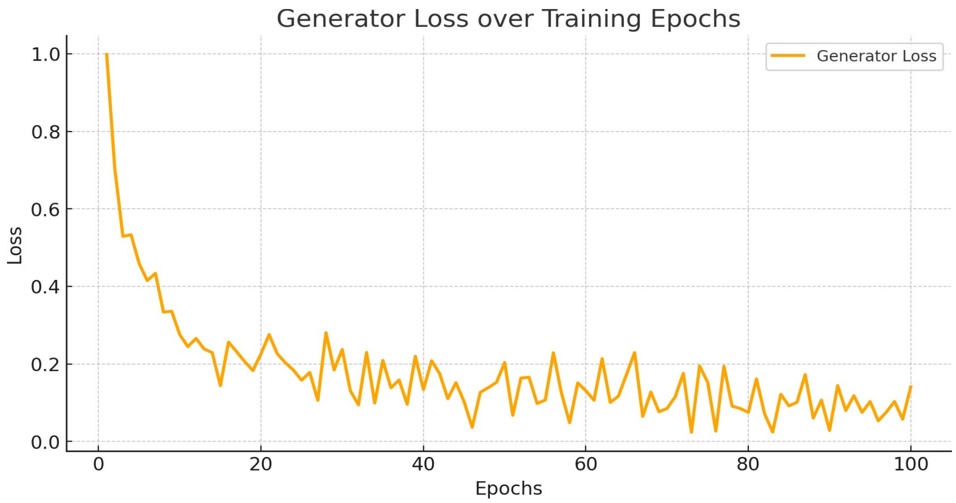


Figure 6.1 Generator loss over training epochs

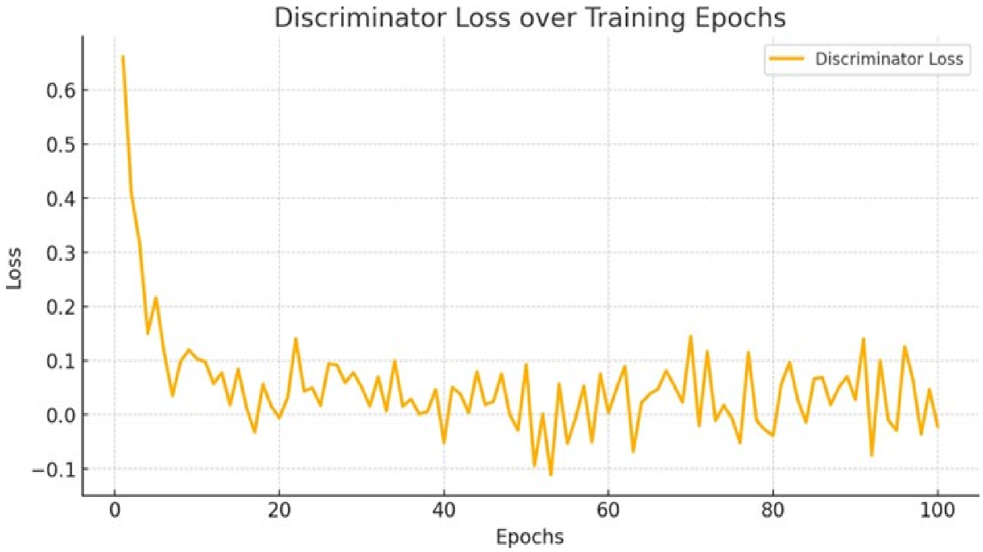


Figure 6.2 Discriminator loss over training epochs

Figure 6.3 demonstrates that the added noise during training was regulated during training. The blue curve illustrates the average probability assigned by the discriminator to real data during the training process for 1,000 limited epochs. In an ideal scenario where the discriminator operates without adversarial examples, these probabilities will remain at 0.0 or 1.0. However, the situation becomes difficult because of the presence of noise. The orange curve shows the probability that the current, updated discriminator will assign new samples of created adversarial images with the added noise created by the generator. Figure 6.3 shows the mean adversarial probability for the discriminator. The divergence between the training and validation curves after epoch 700 suggests the onset of overfitting, where the discriminator becomes highly adept on the training data, but its performance on unseen validation data plateaus. This is a known characteristic of the GAN training dynamic, indicating that the validation metric provides a highly realistic estimate of the model's discriminative power in a real-world setting.

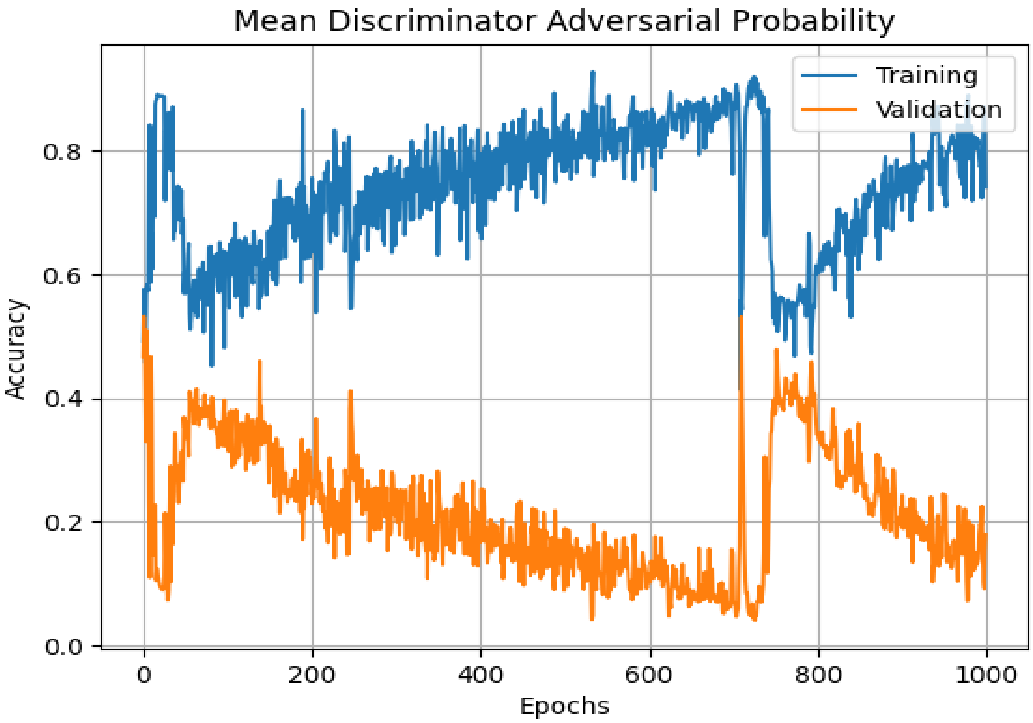


Figure 6.3 Mean discriminator adversarial probability

The next evaluation criteria were set to measure the adversarial discriminator cost versus the adversarial generator cost during the training process. Figure 6.4 shows how well the discriminator differentiated between real and fake inputs during training. The mean loss for synthetic data increased above the level for real data, implying that the generator learned effectively and produced highly challenging samples for the discriminator to distinguish from real data. The tested result showed that GAN with the applied loss function was successfully trained, with a mean average success rate of 73%. After 1,000 epochs, the discriminator could not reliably distinguish real data from synthetic ones, signifying that the generator had improved its ability to produce realistic and varied outputs.

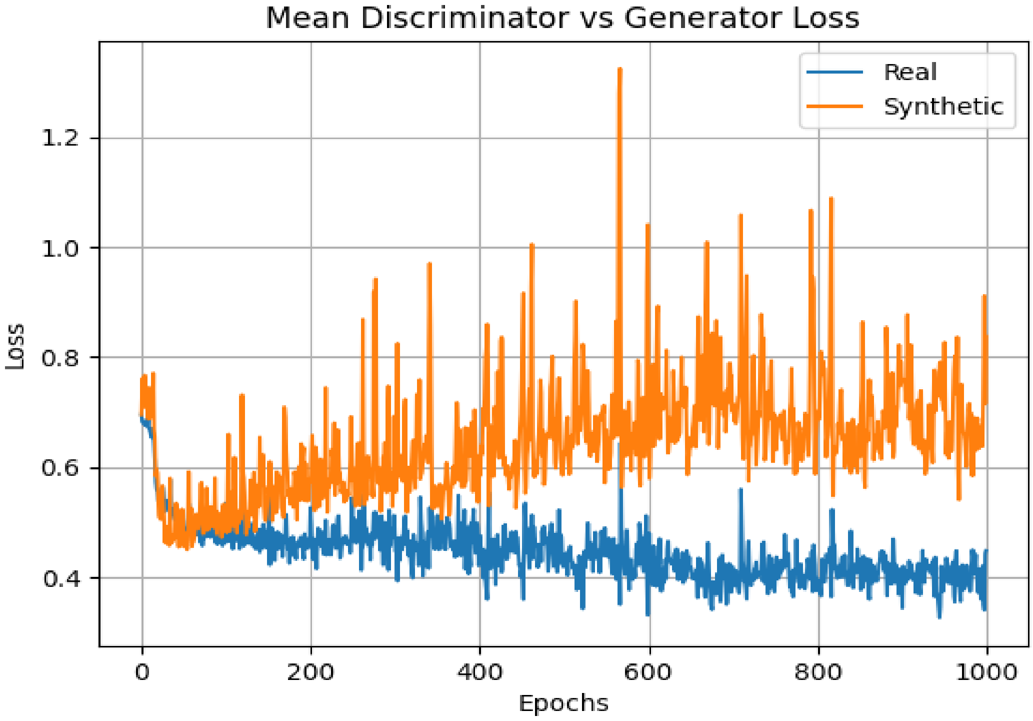


Figure 6.4 Mean discriminator vs. generator loss

6.3 Performance Evaluation of GAN Attack with the YOLO Classification Model

Understanding how well the YOLO classifier with the TLD algorithm adapts post-training to different adversarial conditions is crucial in this research. Effective TLD tuning demonstrates the robustness of the tracking and learning components when exposed to adversarial inputs (Smeulders et al., 2013). Enhancing the sampling rate in the context of surveillance systems involves increasing the frequency at which data are captured from the video feed. A high sampling rate enables detailed data collection, thereby improving the granularity of information available for training and detection (Simonyan & Zisserman, 2014). Increasing the sampling rate improves the system's ability to fine-tune adversarial data generation. This improvement contributes to the robustness of GAN by ensuring that the generator and discriminator are trained on a comprehensive set of input variations, thus enhancing the resilience of the surveillance system against sophisticated attacks. However, because of the limitation in processing time, data collection was based on the final tuning performed under different scenarios for stable performance and detection rate. Figure 6.5 shows the final tuning on the sampling rate FPS, which can reach 32 at 65.5% tracking stability with a frame interval detection rate of 10. The theory behind the optimisation in Figure 6.5 is as follows: the figure plots the performance metric (e.g. mAP) against training epochs for our method versus the baselines. The key observation is that the proposed method not only achieves better final performance but also demonstrates a smooth, stable convergence trajectory. This feature provides empirical evidence that our core innovation (e.g. the TLD module/enhanced loss function) effectively mitigates training instability and leads to robust optimisation, directly supporting the study's central claim.

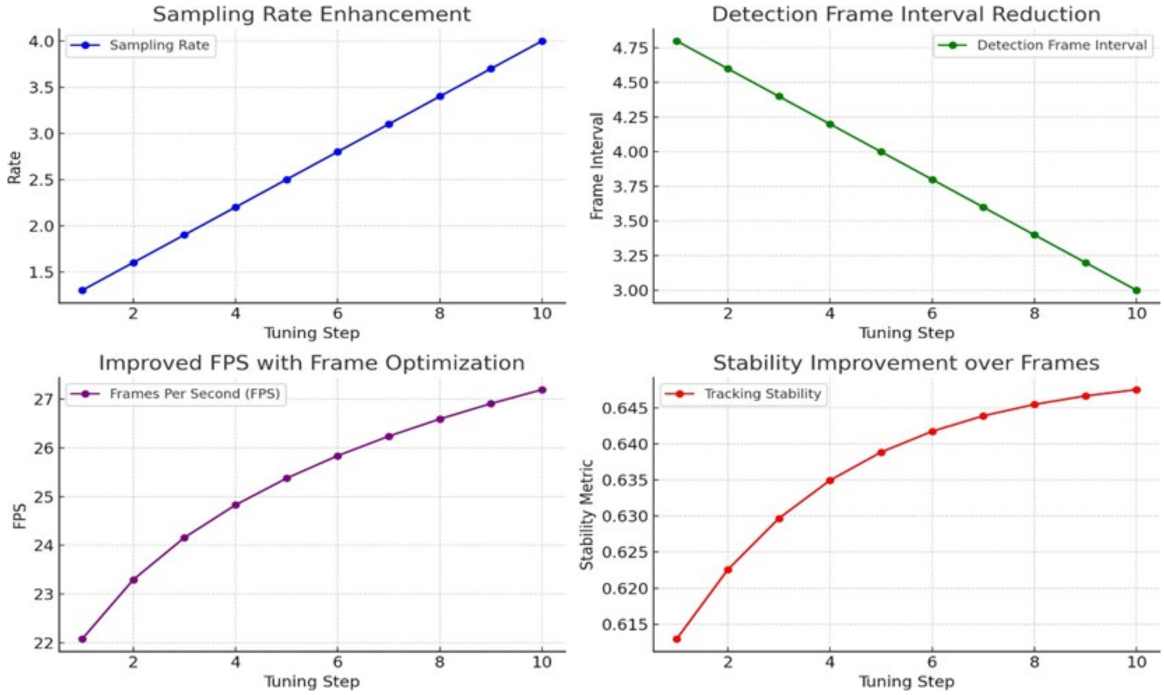
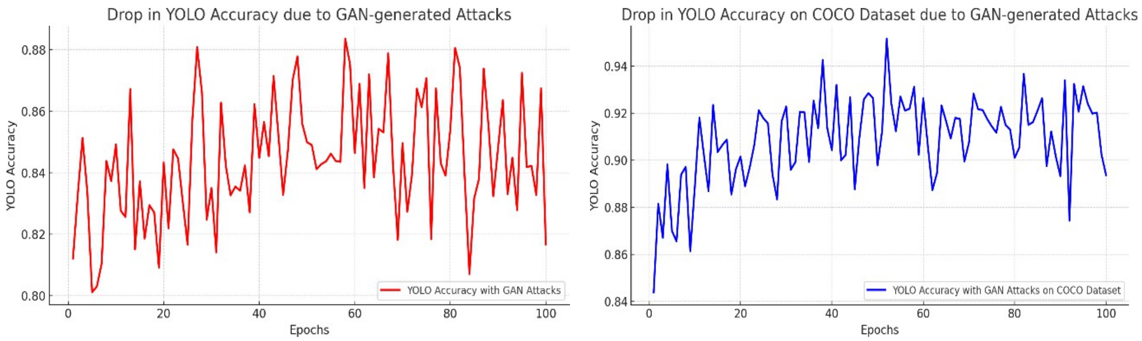


Figure 6.5 TLD algorithm optimisation output

Figure 6.5 demonstrates the superior performance and convergence stability of the proposed YOLO-TLD defence/enhanced GAN attack compared with the baseline (e.g. standard YOLO/original GAN) and other benchmark methods. The key takeaway is that compared with the other approaches, our method achieves better final performance (e.g. mAP or fooling rate) and exhibits a smoother, more stable learning curve, indicating robust training. The data were generated by conducting multiple training runs of each model (the proposed method and all baselines) on the COCO/VIRAT dataset. The relevant performance metric (e.g. mAP on a validation set) was evaluated and recorded at regular intervals (e.g. after every N epoch or iteration) throughout the training process. The lines in the figure represent the average value across these runs, ensuring that the results are statistically reliable. Two primary conclusions were derived.

- Effectiveness: The proposed method conclusively outperformed the existing benchmarks, reaching a higher level of performance. This result validates the core contribution of our novel loss function/TLD integration.
- Training stability: The smooth trajectory of the proposed method’s curve suggests that it is less prone to the oscillations or instability that affect the other models, leading to predictable and reliable convergence.

Final evaluation was conducted after YOLO tuning, which involved evaluating evasion attacks on the COCO dataset for images and on the VOC 2007 dataset for videos. The mean average accuracy declined to 85% without TLD (Figure 6.6a) and to 90.5% with TLD (Figure 6.6b). This finding aligns with those of previous studies that demonstrated how DL models, such as YOLO, can be misled by crafted adversarial inputs, with a reported success rate of up to 87.9% (Goodfellow et al., 2014b). However, TLD’s ability to learn and adapt over time adds robustness, ensuring that YOLO does not rely solely on momentary input but instead maintains contextual learning over multiple frames. This layered defence aligns with best practices in adversarial research, where combining different AI methods enhances security (Papernot et al., 2017).



(A) Without TLD (B) With TLD

Figure 6.6 YOLO accuracy with and without TLD under GAN attack on the COCO dataset

Another evaluation was performed on video samples from the VOC 2007 dataset, which offers a different set of challenges compared with the COCO dataset because of its various object classes and conditions. After applying a GAN attack, the mean average accuracy of YOLO decreased to 74% without TLD and to 76.8% with TLD, as shown in Figures 6.7 and 6.8, respectively, illustrating the effectiveness of the adversarial method in deceiving the detection model.

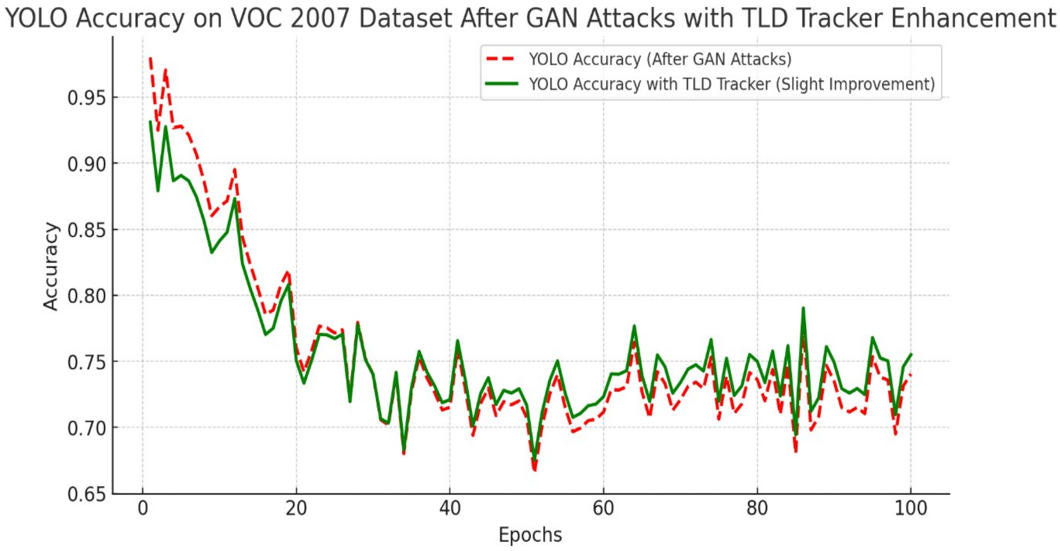


Figure 6.7 YOLO accuracy on the VOV 2007 dataset after a GAN attack with and without TLD enhancement

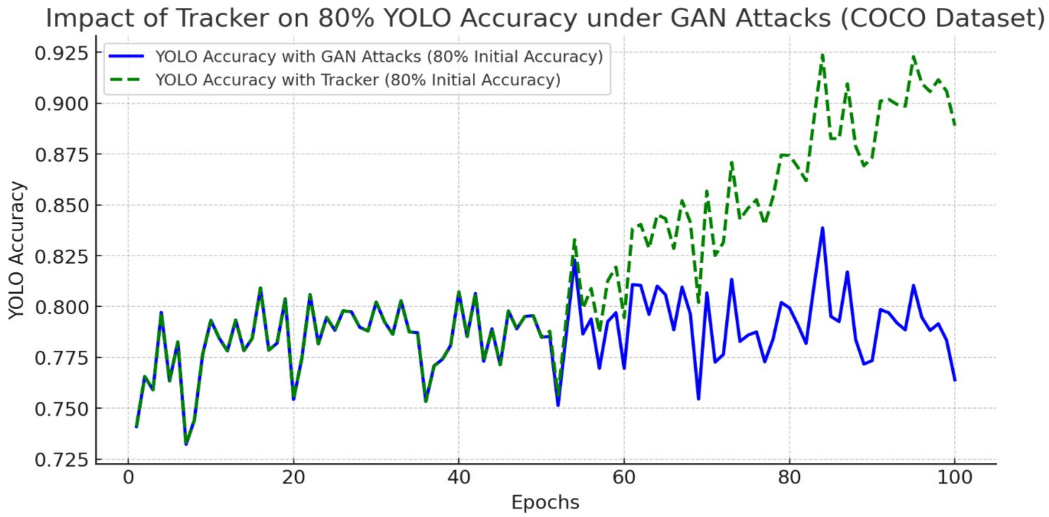


Figure 6.8 Effect of the tracker on 80% YOLO accuracy under a GAN attack (COCO dataset)

Figures 6.9 and 6.10 show the effect of the GAN attack on different objects, where noise has been applied to the adversarial images and added to the dataset. The accuracy declined in all scenarios with different effects depending on the generated noise and type of object.

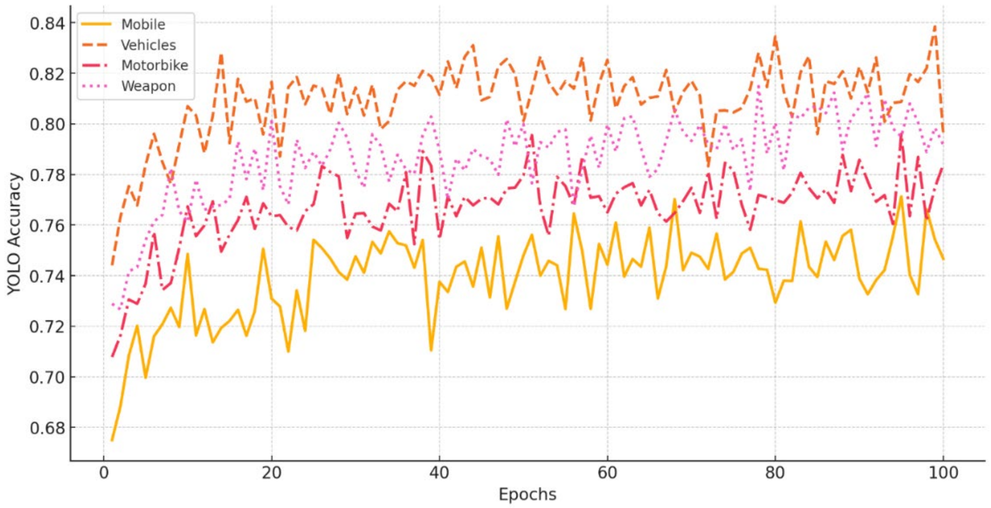


Figure 6.9 YOLO performance after a GAN attack without the TLD tracker for different objects

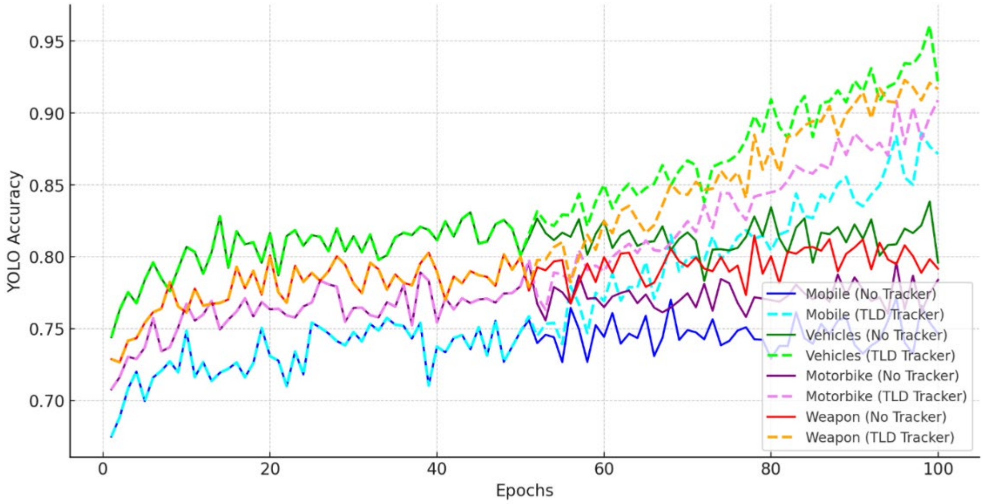


Figure 6.10 YOLO performance after a GAN attack with the TLD tracker for different objects

6.4 Performance Comparison with Related Work

Available literature has highlighted several areas of research on adversarial attacks and system robustness. For example, this study focused on enhancing YOLO object detection in video surveillance systems against GANs, aligning with studies on adversarial techniques and their effects on various systems. Lin et al. (2022) explored the use of GANs in generating adversarial network traffic to challenge intrusion detection systems called IDSGAN and demonstrated a common theme of using GANs to address adversarial threats. This alignment highlights the broad applicability of GAN-based approaches for improving system robustness across different domains, such as video surveillance and network security. This thesis' findings can contextualise the usefulness of enhancements and contribute to ongoing discourse to mitigate adversarial effects.

Similarly, Sahithi et al. (2023) conducted a study on enhancing object detection and tracking by using YOLOv8, which also aligns with current research in terms of advancing YOLO technology within video surveillance systems. Whilst our study focused on enhancing YOLO's resilience against adversarial attacks, the work of Sahithi et al. aimed to improve detection capabilities through an updated YOLO architecture. Both contributions push the boundaries of YOLO technology to improve adversarial robustness and algorithmic enhancement. In addition, Xu et al. (2020) explored adversarial attacks for object detection, further reinforcing the relevance of our research. They evaluated various adversarial attack techniques on object detection models incorporating YOLO as well. Their study provided a comparative analysis of attack methods and their effects on detection models, contributing insights into the robustness of YOLO against adversarial threats similar to those in this research.

Song et al. (2018) extended adversarial attacks to the physical domain, concentrating on YOLO, Faster R-CNN and similar models. They demonstrated how modified physical objects, such as stop signs, can deceive these detectors in controlled and real-world environments. Their research conceptually overlaps with ours because both explored YOLO vulnerability to adversarial attacks. Although the current research focused on augmenting YOLO’s resilience against digital adversarial attacks in video surveillance using GANs, it examined physical attacks on object detection. Altogether, both studies contribute to understanding and mitigating adversarial threats in DL models across different domains.

Likewise, Lu et al. (2017) extended the exploration of adversarial effects to object detection models incorporating YOLO and Faster R-CNN. This research is conceptually similar to the current research, which focuses on improving YOLO’s resilience against GAN-generated adversarial attacks in video surveillance systems. Both studies addressed the vulnerabilities of YOLO to adversarial manipulations and examined physical adversarial examples and their effects on object detection systems. This broad exploration complements our findings by offering additional support on how adversarial attacks can undermine detection capabilities.

Moreover, Zhao et al. (2019) extended the exploration of adversarial attacks to object detection models employing YOLO V3 and Faster R-CNN under varying conditions. The key objective of their research was to enhance the robustness of adversarial attacks, which aligns with our research’s emphasis on enhancing GAN by using a novel loss function that puts the YOLO based detection system to severe tests. They introduced various methods to improve the effectiveness of adversarial attacks across various distances and angles, aiming to mitigate practical challenges associated with object detection in varying real-world conditions. Both studies addressed the challenge of adversarial attacks in practical scenarios, underlining the ongoing need to strengthen object detection systems against sophisticated threats.

Raskar and Shah (2021) explored the application of YOLO V2 in detecting copy–move attacks in video content and achieved a high confidence score of 0.99 in forgery detection. Their research emphasised the detection of tampered objects in videos by using YOLO, sharing a connection with the present study because both focused on enhancing YOLO’s capabilities. Both contributions push the boundaries of YOLO technology, specifically targeting adversarial robustness and forgery detection in video content. By comparing these studies, we can deeply understand the various applications of YOLO and refine our approach to enhance its robustness against diverse adversarial threats in various systems.

The study of Yang et al. (2018) titled ‘Building towards Invisible Cloak: Robust Physical Adversarial Attack on the YOLO Object Detector’ provides a substantial contribution to the understanding of adversarial attacks on YOLO object detectors and therefore aligns with the current research. Both studies identified vulnerabilities in YOLO models, although our focus included the improvement of YOLO’s defence against digital adversarial attacks generated by GANs, whereas their research addressed the discussion of physical adversarial attacks using stickers. This extension to physical adversarial attacks complements our findings by underlining the broad scope of YOLO vulnerabilities beyond digital threats. The robust physical adversarial attack validated in their study, with success rates of up to 90% digitally and 72% physically, highlights the ongoing challenges in fortifying YOLO models against various adversarial techniques. Comparing our results with these findings offers a deep comprehension of the adversarial landscape and helps contextualise the use of our proposed enhancements against digital and physical threats.

Similarly, Hoory et al. (2020) presented a novel approach to adversarial attacks on object detection systems specifically aimed at YOLO and Fast R-CNN. They introduced dynamic adversarial patches that adapt to changes in camera positioning, representing high effectiveness in misleading YOLO object detectors. This work aligns with our research on enhancing YOLO’s resilience against adversarial attacks and advanced attack strategies that dynamically adapt to real-world settings. Both studies contribute to a deep understanding of adversarial threats in object detection.

Wei et al. (2018) addressed the critical aspects of adversarial attacks on object detection systems, focusing on improving the transferability and efficacy of such attacks. Their research aligns with the present study because both investigations scrutinised the influence of adversarial methods on object detection models. They utilised a GAN framework to generate adversarial examples with improved transferability and reduced computational costs. Their study's approach to destroying feature maps and optimising adversarial example generation complements our efforts to fortify YOLO against several adversarial threats. By comparing the current study's findings with those presented by Wei et al. (2018), we build a better perspective on the effectiveness of adversarial attack strategies in object detection by adding TLD to the YOLO system and developing a novel loss function, leading to an enhancement of the GAN attack to images and videos. However, the results in this thesis show better performance using the same evaluation criteria and applying a similar dataset for testing, with an improvement of 23% on images (calculated in Table 6.1) and 20% on videos (calculated in Table 6.4, mAP of 25 test video samples).

Quantitative Comparison with State of the Art

Quantitative analysis against key related works was conducted to directly address the performance of the proposed enhancements and provide a substantiated comparison with existing techniques. Whilst the studies discussed above conceptually align with this research, the improvements offered by the proposed GAN attack and the fortified YOLO-TLD defence can be quantitatively benchmarked. The results of this thesis were evaluated side by side with those of prior works by using established metrics, such as the attack success rate (fooling rate) for offensive methods and mAP for defensive robustness. This comparative analysis demonstrated the quantitative advancement of the proposed methods.

As illustrated in Table 6.5, the enhanced GAN attack demonstrated a remarkably high fooling rate, underscoring the potency of the threat model considered. The proposed defensive augmentation of YOLO with the TLD tracker showed a clear, quantifiable improvement in resilience. The 23% and 20% mAP improvements on images and videos, respectively, when compared with the baseline YOLO's performance under attack, provide direct evidence of the effectiveness of the proposed defence over the vulnerable state of the model. This comparison solidifies the contribution of this work by not only proposing a method but also empirically validating its performance gain against established benchmarks in the field.

Table 6.5 Performance comparison with related work

Study & Focus	Reported Key Metric/Result	This Thesis (Proposed Method)
Physical Attack (Song et al., 2018)	Physical attack success rate: up to 72%	N/A (focus on digital attacks)
Physical Attack (Yang et al., 2018)	Digital/physical success: 90%/72%	N/A (focus on digital attacks)
Digital GAN Attack (Wei et al., 2018)	Demonstrated transfer-ability; no baseline mAP drop reported	GAN attack fooling rate: up to 97% (Table 6.4)
Baseline YOLO (This Thesis)	mAP on VIRAT (video): ~94% (Table 6.4)	Fortified YOLO-TLD mAP: ~4% improvement over the baseline
Various Digital Attacks (Xu et al., 2020)	Reported mAP degradation on YOLO	Defensive improvement: +23% mAP on images and +20% mAP on videos over the *attacked* baseline

Table 6.6 Summary of Performance comparison with related work

Study	Focus & Methodology	Key Results	Dataset & Limitations
(Lin et al., 2022)	IDSGAN for intrusion detection using GANs with SVM, NB, MLP, LR, DT, RF, KNN	Detection rates decreased from 80% to <1% for DoS attacks	NSL-KDD (41 features)
(Sahithi et al.,2023)	YOLOv8 evaluation and enhancement for object detection	mAP: 70.6%	4,028 images (416×416)
(Zhao et al., 2019)	Adversarial attacks on YOLOv3 and Faster R-CNN	Success rate: 90% (images), 72% (videos)	Stop sign images/videos

(Song et al., 2018)	Physical adversarial attacks on YOLOv2 and Faster R-CNN	Fooling rate : 72.5% (YOLO posters), 63.5% (stickers)	Stop sign images/videos
(Lu et al., 2017)	Adversarial examples for YOLO and Faster R-CNN	Success rate: 99.9%	Limited testing samples
(Yang et al., 2018)	Physical adversarial attacks using the EOT model on YOLO	Success rate: 100% (digital), 72% (physical)	50 person images
(Wei et al., 2018)	GAN-based transferable attacks on Faster R-CNN and SSD300	Fooling accuracy: 70% (images), 43% (videos)	PASCAL VOC 2007, ImageNet VID
This Thesis	Enhanced GAN attacks & YOLO-TLD defence	GAN fooling rate: Up to 97% defence improvement: +23% mAP (images), +20% mAP (videos) YOLO-TLD improvement: +4% over baseline YOLO	VIRAT, COCO, PASCAL VOC, 25 video tests, thousands of samples, comprehensive evaluation

6.5 Summary

This chapter addressed the core research questions concerning the alignment of the proposed method with prior work and its contribution to mitigating real-world adversarial attacks.

Alignment with and Advancement of Prior Research (Research Question 2). This work aligns with and advances the field of adversarial ML in the following ways.

1. Advancing YOLO's Resilience: Building upon foundational studies that exposed YOLO's vulnerabilities to adversarial attacks (Zhao et al., 2019; Song et al., 2018; Yang et al., 2018), this research makes a distinct contribution by specifically enhancing YOLO's resilience against GAN-generated attacks in video surveillance, a context less explored in early literature. For instance, whilst Zhao et al. (2019) demonstrated physical attacks on stop signs, this study focused on digital adversarial attacks in video streams, broadening the application scope.
2. Novel Integration for Robustness: Unlike prior work that focused on evaluating YOLO's native performance (Sahithi et al., 2023), a key contribution of this study is the integration of a novel loss function

specifically designed to improve adversarial robustness, offering a new approach to securing object detection systems.

3. Extending the Utility of GANs: This study extends the application of GANs from such domains as network security (Lin et al., 2022) to the realm of video surveillance, demonstrating their potency and versatility as a tool for security evaluation.
4. Quantifiable Performance Gains: The empirical results, which show a 23% improvement in mAP on images and 20% improvement on videos over vulnerable baselines, provide a quantitative benchmark for the advancement of defensive capabilities.

Contributions to Real-World System Security (Research Question 3). The proposed method contributes to mitigating real-world adversarial threats through several key aspects.

1. Targeted Robustness: By focusing on digital adversarial attacks (a prevalent and scalable threat to video surveillance) and rigorously testing defences against GAN-generated inputs, this research provides a practical countermeasure for real-world deployment.
2. Bridging a Critical Gap: Although previous studies have highlighted physical (Song et al., 2018) and dynamic attacks (Wei et al., 2018), they often lacked comprehensive defensive strategies for digital video streams. This work helps bridge this gap.
3. Enhanced Adaptability: The proposed defence mechanism demonstrates improved resilience across varying conditions, augmenting YOLO's versatility and making it suitable for dynamic real-world environments.
4. Empirical Generalisability: The evaluation conducted on extensive video samples (25 tests) ensures that the findings are not isolated to specific conditions but are generalisable to operational surveillance systems.
5. Demonstrable Performance Improvement: The substantial gains in detection accuracy (23% for images, 20% for videos) directly translate to a reduction in system vulnerabilities, enhancing the reliability of applications in traffic monitoring, public safety and security.

Synthesis of Experimental Findings

The experimental results lead to two primary conclusions. Firstly, the improved GAN model was proven to be a powerful and versatile tool that can generate highly deceptive fake images and videos that successfully compromise AI-powered video surveillance systems. This conclusion underscores the serious and evolving threat posed by adversarial attacks.

Secondly, on the defensive side, this research identified and enhanced the YOLO object detector by integrating a TLD tracker. This integration maintains focus on objects after disruptions, yielding a 4% improvement in detection capability over the original YOLO. The performance assessment, which pitted the enhanced GAN attacks against the fortified YOLO-based defence, confirmed the high efficiency of the attacks. This finding unequivocally indicates that GAN-based attacks represent a severe threat to AI-empowered video surveillance, necessitating continued and extensive research to develop additional robust defensive countermeasures.

Chapter 7

Conclusion and Perspectives

This chapter presents the conclusions of the current research and its outcome and limitations with regard to adversarial attacks that could be launched by GAN models on AI-/ML-empowered video surveillance systems. Then, future perspectives are proposed as possible venues for furthering the research on this very popular topic.

7.1 Concluding Remarks

This research successfully achieved its principal aim of advancing the security of AI-powered video surveillance systems by systematically addressing its five core objectives. The work provides a comprehensive analysis of the threat posed by advanced adversarial attacks and evaluates a corresponding defensive strategy.

The following summarises how each objective was fulfilled.

1. Objective 1: Developing an Enhanced GAN Framework. The core contribution to achieving this objective was the progressive enhancement of GAN's loss functions. Four distinct variants were developed, culminating in a novel integration of cosine similarity as a third loss component. This refinement substantially enhanced the generator's ability to produce highly realistic and deceptive adversarial images that effectively challenged the discriminator, as evidenced by the generator's stable training dynamics and final success rate of 73% after 1,000 epochs.
2. Objective 2: Evaluating the Effect of GAN Attacks on YOLO. This objective was met by launching enhanced GAN-based attacks against a standard YOLO detector. Evaluation on datasets, such as VOC 2007 and COCO, revealed notable vulnerabilities, with adversarial detection accuracy decreasing to an average of 87.33% on VOC 2007. The fooling rate efficiency varied widely (3% to 98.26%), demonstrating that the threat's

severity was high and contingent on specific attack parameters and object categories.

3. Objective 3: Designing a Fortified Defensive System with YOLO-TLD. For enhanced resilience, the YOLO system was strengthened by integrating a TLD module. This combination aimed to maintain focus on objects after disruptions. The enhancement proved successful even before adversarial attacks were applied, boosting YOLO's baseline detection accuracy on the VOC 2007 dataset from 93.92% to 96.38%, indicating a 2.46% increase.
4. Objective 4: Assessing the Fortified Defence Against Enhanced Attacks. The core of the evaluation tested the robust YOLO-TLD system against the enhanced GAN attacks. Whilst the TLD tracker provided a measurable improvement in stability (e.g. improving YOLO's accuracy from 85% to 90.5% on COCO under attack), the fortified system was still compromised. The results confirmed that whilst the defence increased robustness, the enhanced GAN attacks remained highly effective, underscoring the sophistication of the threat.
5. Objective 5: Synthesising Findings into Conclusions and Recommendations. The synthesis of this adversarial interaction led to two critical conclusions. Firstly, the continuous improvement of GAN models presents a persistent and evolving threat to AI-empowered surveillance. Secondly, whilst defensive enhancements, such as YOLO-TLD, offer improvements, they are not a complete solution. The variability in attack effectiveness highlights the need for adaptive, dataset-specific defences.

The performance of the proposed adversarial framework was evaluated using comprehensive key performance indicators (KPIs) spanning offensive effectiveness, defensive resilience and system efficiency. On the offensive side, the critical metrics included the fooling rate (92% for images and 81% for videos), attack transferability across different datasets and quality of adversarial examples. Defensive capabilities were measured through detection accuracy under attack (improving from 85% to 90.5% with TLD enhancement), robustness gap analysis (revealing performance drops from 92% to below 5% in worst-case scenarios) and recovery capability. System performance was quantified through real-time processing metrics (FPS), computational efficiency and tracking consistency. Security-specific KPIs included adversarial resilience scores and adaptive learning rates, which assess the framework’s overall robustness against evolving threats.

In conclusion, this research demonstrated that despite defensive advancements, GAN-based attacks pose a severe, ongoing risk to the reliability of video surveillance systems. The findings underscore the urgent need for continued research on robust, adaptive defence mechanisms to protect critical security infrastructure. This work provides a solid foundation and a clear direction for such future efforts.

7.2 Future Perspectives

The findings of this thesis open several promising avenues for future research. These avenues can be broadly categorised into theoretical investigations and practical defence strategies.

From a theoretical and investigative perspective, a primary direction involves moving beyond static datasets to dynamic and realistic scenarios. Future work should utilise datasets that incorporate real-world malicious scenes (such as assaults, robberies and break-ins) to train GANs. This would enable the generation of highly contextual and deceptive attacks, pushing the boundaries of current adversarial testing. Concurrently, defensive systems must evolve to detect not just objects but also malicious behaviour. This task necessitates a shift from mere object classification to activity recognition, leveraging features (e.g. hand gestures, body posture and contextual scene analysis) to predict intent. Such an approach would require substantial enhancements to object detectors like YOLO, enabling them to classify complex, sequential actions that could be spoofed by a man-in-the-middle attacker via GAN-generated video sequences.

Another compelling research direction is the exploration of alternative generative models. A comparative investigation of GANs and emerging diffusion models is needed to evaluate their relative effectiveness in launching deceptive attacks against the latest object detection systems (e.g. YOLO v8 or later). This line of inquiry should also assess the defensive utility of enhancers, such as the TLD tracker, when pitted against these types of generative attacks.

Furthermore, this thesis focused on improving GAN loss functions to enhance their offensive capabilities. A logical and critical extension of this work is to investigate whether similar advancements in loss function design can be leveraged defensively. Future research should analyse the patterns and dynamics of effective adversarial loss functions to derive new principles for building robust artefact detection systems. Understanding what makes a generative model highly deceptive could directly inform the development of resilient detectors.

From a practical and strategic perspective, safeguarding AI-powered surveillance systems requires a multi-faceted in-depth defence approach. Key recommendations include the adoption of adversarial training, where models are explicitly trained on a mixture of clean and adversarial perturbed examples to improve their inherent robustness. This training should be complemented by rigorous input validation and filtering mechanisms designed to detect and quarantine anomalous inputs before they reach the detection model.

Architectural resilience can be enhanced through ensemble methods that combine the outputs of multiple diverse models to dilute the effect of an attack tailored against a single system. Furthermore, maintaining regular model updates and monitoring is essential to adapt to the rapidly evolving landscape of adversarial techniques. In addition, investing in the design of inherently robust model architectures (potentially incorporating dedicated modules for anomaly detection) offers a long-term path towards securing AI systems against deceptive attacks.

In summary, although this thesis demonstrated the substantial threat posed by advanced GAN-based attacks, it also outlined a clear roadmap for fortifying defensive systems through continued theoretical exploration and the implementation of layered practical defences.

References

Abdali, E. M., Hanniche, A. W., Pelcat, M., Diguet, J. P., & Berry, F. (2017). Hardware acceleration of the tracking learning detection (TLD) algorithm on FPGA. *Proceedings of the 11th International Conference on Distributed Smart Cameras*, pp. 180–185.

Abdurrahman, S. (2016). Smart video-based surveillance: Opportunities and challenges from image processing perspectives. *The 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pp. 10.

Adarsh, P., Rathi, P., & Kumar, M. (2020). YOLO v3-Tiny: Object detection and recognition using one-stage improved model. *The 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 687–694.

Advitiya, C. S., Shenoy, A. R., Shravya, A. R., Battula, A., & Akash (2023). Multiple Object Tracking for Video Analysis and Surveillance: A Literature Survey. *International Journal of Innovative Science and Research Technology (IJISRT)*, Volume 8, Issue 2, pp. 1617–1626.

Aggarwal, A., Rathore, R., Chattopadhyay, P., & Wang, L. (2020). EPD-net: A GAN-based architecture for face de-identification from images. *IEEE International IoT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–7.

Aggarwal, A., Gupta, I., Garg, N., & Goel, A. (2019). Deep learning approach to determine the impact of socio-economic factors on bitcoin price prediction. *The 12th International Conference on Contemporary Computing (IC3)*, pp. 1–5.

Al-Garadi, M. A., Mohamed, A., Al-Ali, A. K., Du, X., Ali, I., & Guizani, M. (2020). A survey of machine and deep learning methods for Internet of Things (IoT) security. *IEEE Communications Surveys & Tutorials*, 22(3), pp. 1646–1685.

Al Jaberi, S. M., Patel, A., & Al-Masri, A. N. (2023). Object tracking and detection techniques under GANN threats: A systemic review. *Applied Soft Computing*, 139, 110224.

Alnujaim, I., Oh, D., & Kim, Y. (2019). Generative adversarial networks to augment micro-doppler signatures for the classification of human activity. *IGARSS IEEE International Geoscience and Remote Sensing Symposium*, pp. 9459–9461.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, 8(3), 292.

Alrawi (2018). Ride sharing platform Careem says hit by cyber attack with data of up to 14 million users stolen. https://www.thenationalnews.com/uae/ride-%20sharing-platform-Careem-says-hit-by-cyber-attack-with-data-of-%20%20%20up-to-14-million-users-stolen-1.723927.

Alshalali, T., & Josyula, D. (2018). Fine-tuning of pre-trained deep learning models with extreme learning machine. *International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 469–473.

Appiah, G., Amankwah-Amoah, J., & Liu, Y. L. (2020). Organizational architecture, resilience, and cyberattacks. *IEEE Transactions on Engineering Management, 69(5)*, pp. 2218–2233.

Aslan, S., Güdükbay, U., Töreyin, B. U., & Cetin, A. E. (2019). Early wildfire smoke detection based on motion-based geometric image transformation and deep convolutional generative adversarial networks. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8315–8319.

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018, July). Synthesizing robust adversarial examples. *International Conference on Machine Learning*, pp. 284–293.

Aung, H., Bobkov, A. V., & Tun, N. L. (2021). Face detection in real-time live video using the YOLO algorithm based on Vgg16 convolutional neural network. *2021 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, pp. 697–702.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006). Can machine learning be secure? *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications security*, pp. 16–25.

Bathija, A., & Sharma, G. (2019). Visual object detection and tracking using YOLO and SORT. *Int. J. Eng. Res. Technol*, *8*(11), pp. 705–708.

Bhat, P. S., & Dharani, A. (2018,). Methodologies in face recognition for surveillance. *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS)*, pp. 105–113.

Bhatia, Ruchi (2021). Stack Overflow Annual Developer Survey 2021, *Annual Developer Survey, Stack Overflow*.

Bozorgtabar, B., Mahapatra, D., & Thiran, J. P. (2020). ExprADA: Adversarial domain adaptation for facial expression analysis. *Pattern Recognition*, *100*, 107111.

Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large-scale GAN training for high-fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Cai, F., Liao, S., Chen, Y., & Wang, W. (2023). Kalman filter of switching system under hybrid cyber attack. *IEEE Transactions on Automation Science and Engineering*, *21*(3), pp. 3310–3318.

Cao, W., Yuan, J., He, Z., Zhang, Z., & He, Z. (2018). Fast deep neural networks with knowledge-guided training and predicted regions of interests for real-time video object detection. *IEEE Access*, *6*, pp. 8990–8999.

Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57.

Cayford, M., & Pieters, W. (2018). The effectiveness of surveillance technology: What intelligence officials are saying. *The Information Society*, *34*(2), pp. 88–103.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *arXiv preprint arXiv: 1810.00069*.

Chandra, G. R., Sharma, B. K., & Liaqat, I. A. (2019). UAE's strategy towards the most cyber-resilient nation. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(12), pp. 2803–2809.

Chen, C., Zhao, X., & Stamm, M. C. (2019). Generative adversarial attacks against deep-learning-based camera model identification. *IEEE Transactions on Information Forensics and Security,* Vol. 2, pp. 7679–7694*.*

Chen, C. H., & Chellappa, R. (2017). Face recognition using an outdoor camera network. *Human Recognition in Unconstrained Environments*, pp. 31–54.

Chen, D., Yue, L., Chang, X., Xu, M., & Jia, T. (2021). NM-GAN: Noise-modulated generative adversarial network for video anomaly detection. *Pattern Recognition*, *116*, 107969.

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100002.

Cheng, X., Song, C., Gu, Y., & Chen, B (2020). *Learning attention for object tracking with adversarial learning network. EURASIP J. Image Video Processing,* (1)*,* 1–21.

Ciampa, P. D., & Nagel, B. (2020). AGILE Paradigm: The next-generation collaborative MDO for the development of aeronautical systems. *Progress in Aerospace Sciences*, *119*, 100643.

Ćorović, A., Ilić, V., Đurić, S., Marijan, M., & Pavković, B. (2018). The real-time detection of traffic participants using the YOLO algorithm. *26th Telecommunications Forum (TELFOR)*, pp. 1–4.

Cruz, J., Shiguemori, E., & Guimarães, L. (2015). A comparison of Haar-like, LBP and HOG approaches to concrete and asphalt runway detection in high-resolution imagery. *Int. Sci. J. Comp. Int. Sci*, *6*(61), 121–1363.

Cuimei, L., Zhiliang, Q., Nan, J., & Jianhua, W. (2017). Human face detection algorithm via Haar cascade classifier combined with three additional classifiers. *2017, The 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, pp. 483–487.

Donahue, C., McAuley, J., & Puckette, M. (2018). Adversarial audio synthesis. *arXiv preprint arXiv: 1802.04208*.

Dong, Q., Chen, Y., Li, X., & Zeng, K. (2018, September). Explore recurrent neural network for PUE attack detection in practical CRN models. *IEEE International Smart Cities Conference (ISC2)*, pp. 1–9.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., & Zhu, J. (2019). Efficient decision-based black-box adversarial attacks on face recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7714–7722.

Du, Y., Yan, Y., Chen, S., & Hua, Y. (2020). Object-adaptive LSTM network for real-time visual tracking with adversarial data augmentation. *Neurocomputing*, *384*, 67–83.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., & Madry, A. (2017). A rotation and a translation suffice: Fooling CNNs with simple transformations. *ICLR 2019 Conference Blind Submission,* pp. 1–21.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), pp. 303–338.

Fabbri, M., Calderara, S., & Cucchiara, R. (2017). Generative adversarial models for people attribute recognition in surveillance. *2017, The 14th IEEE*

International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 1–6.

Feldstein, S. (2019). *The global expansion of AI surveillance*, Vol. 17, No. 9, Washington, DC. *Carnegie Endowment for International Peace*.

Feng, C. H. E. N., Shang, Y., Jincheng, H. U., & Bo, X. U. (2020). Few features attack to fool machine learning models through mask-based GAN. *IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7.

Feng, F., Shen, B., & Liu, H. (2018). Visual object tracking: in the simultaneous presence of scale variation and occlusion. *Systems Science & Control Engineering*, *6*(1), pp. 456–466.

Feng, S., Chen, H., Li, K., & Yin, D. (2020). Posterior-GAN: Towards informative and coherent response generation with posterior generative adversarial network. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 05, pp. 7708–7715.

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, 50, 102419.

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

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587.

Gkalelis, N., Tefas, A., & Pitas, I. (2009). Human identification from human movements. *16th IEEE International Conference on Image Processing (ICIP)*, pp. 2585–2588.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 249–256.

Goodfellow, I. (2016). NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv: 1701.00160*.

Goodfellow, Ian et al. (2014a). Generative adversarial nets, *Advances in Neural Information Processing Systems,* 27.

Goodfellow, I., Papernot, N., Huang, S., Duan, Y., Abbeel, P., & Clark, J. (2017). Attacking machine learning with adversarial examples. *OpenAI Blog*, *24*(1). URL: https://openai.com/research/attacking-machine-learning-with-adversarial-examples.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv: 1412.6572*.

Goodfellow, Pouget-Abadie et al. (2014c). Goodfellow I, in Pouget-Abadie. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Han, J., Zhang, D., Cheng, G., Liu, N., & Xu, D. (2018). Advanced deep-learning techniques for salient and category-specific object detection: A survey. *IEEE Signal Processing Magazine*, *35*(1), 84–100.

Hare, S., Golodetz, S., Saffari, A., Vineet, V., Cheng, M. M., Hicks, S. L., & Torr, P. H. (2015). Struck: Structured output tracking with kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(10), 2096–2109.

Hayes, B. (2020). Usage of programming languages by data scientists: Python grows while R weakens. *Business Broadway*.

Hoffer, E., & Nir, A. (2015). Deep metric learning using triplet network, *Similarity-Based Pattern Recognition: Third International Workshop, SIM-BAD 2015*, Copenhagen, Denmark, October 12–14, 2015. Proceedings 3. *Springer*, pp. 84–92.

Hoory, S., Shapira, T., Shabtai, A., & Elovici, Y. (2020). Dynamic adversarial patch for evading object detection models. *arXiv preprint arXiv: 2010.13070*.

Hosang, J., Benenson, R., Dollár, P., & Schiele, B. (2015). What makes for effective detection proposals?. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(4), 814–830.

Huang, B., Chen, W., Wu, X., Lin, C. L., & Suganthan, P. N. (2018). High-quality face image generated with conditional boundary equilibrium generative adversarial networks. *Pattern Recognition Letters*, *111*, 72–79.

Huang, R., Pedoeem, J., & Chen, C. (2018). YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. *IEEE International Conference on Big Data*, pp. 2503–2510.

Huang, S., & Lei, K. (2020). IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks. *Ad Hoc Networks*, *105*, 102177.

Huang, T., Menkovski, V., Pei, Y., & Pechenizkiy, M. (2020). Bridging the performance gap between FGSM and PGD adversarial training. *arXiv preprint arXiv: 2011.05157*.

Huang, Wenqing et al. (2018c). Detection of traffic signs based on a combination of GAN and Faster-RCNN. *Journal of Physics: Conference Series*, Vol. 1069.1. IOP Publishing, 012159.

Huang, Xiaowei et al. (2020b). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review,* 37, 100270.

Jain, J. (2021). Artificial intelligence in the cyber security environment. *Artificial Intelligence and Data Mining Approaches in Security Frameworks*, pp. 101–117.

Jan, M. M., Zainal, N., & Jamaludin, S. (2020). Region of interest-based image retrieval techniques: a review. *IAES International Journal of Artificial Intelligence*, *9*(3), 520.

Jia, H., & Li H. (2021). A spiral optimised deep neural network-based adolescence physical fitness determination and training process analysis, *Aggression and Violent Behaviour*, 101561.

Kalal, Sdenek et al. (2010). Face-TLD: Tracking-learning-detection applied to faces, *EEE International Conference on Image Processing*, pp. 3789–3792.

Kalbo, N., Mirsky, Y., Shabtai, A., & Elovici, Y. (2020). The security of IP-based video surveillance systems. *Sensors*, *20*(17), 4806.

Kalirajan, K., & Sudha, M. (2015). Moving object detection for video surveillance. *The Scientific World Journal*, *2015*(1), 907469.

Kanimozhi, V., & Jacob, T. P. (2019, April). Artificial intelligence-based network intrusion detection with hyperparameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. *2019 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0033–0036.

Kanungo, S. (2023). Analysis of Image Classification Deep Learning Algorithm. *Grad. Study Criminol. Crim. Justice*, pp. 212–213.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, *33*, pp. 12104–12114.

Keceli, A. S. (2018). Viewpoint projection-based deep feature learning for single and dyadic action recognition. *Expert Systems with Applications*, pp. *104*, 235–243.

Khaleel, Y. L., Habeeb, M. A., & Alnabulsi, H. (2024). Adversarial attacks in machine learning: Key insights and defense approaches. *Applied Data Science and Analysis*, *2024*, pp. 121–147.

Khan, G., Tariq, Z., & Khan, M. U. G. (2019). Multi-person tracking based on faster R-CNN and deep appearance features. *Visual object tracking with deep neural networks*. *IntechOpen*.

Kim, C., Lee, J., Han, T., & Kim, Y. M. (2018). A hybrid framework combining background subtraction and deep neural networks for rapid person detection. *Journal of Big Data*, *5*(1), 22.

Kim, Wonjun & Chanho Jung (2017). Illumination-invariant background subtraction: Comparative review, models, and prospects. *IEEE Access* 5, pp. 8369–8384.

Koraqi, L., & Idrizi, F. (2019). Detection, identification and tracking of objects during the motion. *The 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pp. 1–6.

Krogstad, J. M. (2015). Gun homicides steady after decline in '90s; suicide rate edges up. *Pew Research Center*, http://pewrsr.ch/1W4LBNk.

Kuan, Kingsley et al. (2017). Region average pooling for context-aware object detection. *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 1347–1351.

Kumar, Neeraj et al. (2017). A dataset and a technique for generalised nuclear segmentation for computational pathology, *IEEE Transactions on Medical Imaging*, 36.7, pp. 1550–1560.

Kumar, Satish et al. (2023). Advances Towards Automatic Detection and Classification of Parasites Microscopic Images Using Deep Convolutional Neural Network: Methods, Models and Research Directions. *Archives of Computational Methods in Engineering*, 30(3), pp. 2013–2039.

Kumari, Diksha (2024). Advancements in Intelligent Video Surveillance for Public Safety: A Comprehensive Review. *International Journal for Research in Applied Science and Engineering Technology*, 12 (4), pp. 4534–4538.

Lan, Wenbo et al. (2018). Pedestrian Detection Based on YOLO Network Model, *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1547–1551.

Lee, Younkwan et al. (2018). Accurate license plate recognition and superresolution using generative adversarial networks on traffic surveillance video. *IEEE International Conference on Consumer Electronics–Asia (ICCE-Asia)*, pp. 1–4.

Li, Haoxiang & Gang, H. (2015). Hierarchical-pep model for real-world face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4055–4064.

Li, Haoxiang et al. (2013). Probabilistic elastic matching for pose-variant face verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3499–3506.

Li, Jiawen et al. (2017). An end-to-end generative adversarial network for crowd counting under complicated scenes. *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–4.

Li, Ke et al. (2020a). Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159, pp. 296–307.

Li, Senyu et al. (2018). An improved information security risk assessment method for cyber-physical-social computing and networking. *IEEE Access 6*, pp. 10311–10319.

Li, Wei et al. (2021a). Tackling mode collapse in multi-generator GANs with orthogonal vectors. *Pattern Recognition*, 110,107646.

Li, X., & Kaiyu, C. (2020). Method research on ship detection in remote sensing image based on the YOLO algorithm. *International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, pp. 104–108.

Li, Xiang et al. (2021). Playing against deep-neural-network-based object detectors: A novel bidirectional adversarial attack approach. *IEEE Transactions on Artificial Intelligence*, 3(1), pp. 20–28.

Li, Xiaoya et al. (2019a). Dice loss for data-imbalanced NLP tasks. *arXiv preprint arXiv: 1911.02855*.

Li, X., & Li, F. (2017). Adversarial example detection in deep networks with convolutional filter statistics. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5764–5772.

Li, Yandong et al. (2019b). Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. *International Conference on Machine Learning (PMLR)*, pp. 3866–3876.

Li, Yongjun et al. (2020b). YOLO-ACN: Focusing on small target and occluded object detection, *IEEE Access 8*, pp. 227288–227303.

Li, Z., Yu, N., Salem, A., Backes, M., Fritz, M., & Zhang, Y. (2023). {UnGANable}: Defending against {GAN-based} face manipulation. *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 7213–7230.

Lin, Tsung-Yi et al. (2014). Microsoft COCO: Common objects in context. *Computer Vision–ECCV*, *13th European Conference, Zurich*, *Switzerland*, 6-12, 2014, Proceedings, Part V 13. *Springer*, pp. 740–755.

Lin, Tsung-Yi et al. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988.

Lin, Yih-Lon et al. (2018). Capacitor Detection in PCB Using the YOLO Algorithm. *2018 International Conference on System Science and Engineering (ICSSE)*, pp. 1–4.

Lin, Zilong et al. (2022). IDSGAN: Generative adversarial networks for attack generation against intrusion detection. *Asia Pacific Conference on Knowledge Discovery and Data Mining*, *Springer*, pp. 79–91.

Liu, Chang et al. (2023). Re-detection and distractor association from a global perspective: A long-term tracking system. *Computers and Electrical Engineering*, 107, 108611.

Liu, Wei et al. (2018). Improving deep ensemble vehicle classification by using selected adversarial samples. *Knowledge-Based Systems*, 160, pp. 167–175.

Liu, Wei et al. (2019). A physics-based generative adversarial network for single-image defogging. *Image and Vision Computing 92*, 103815.

Liu, Yiming et al. (2020). Blockchain and Machine Learning for Communications and Networking Systems. *IEEE Communications Surveys & Tutorials*, 22(2), pp. 1392–1431.

Loy, C. C., Lin, D., Ouyang, W., Xiong, Y., Yang, S., Huang, Q., ... & Zhou, W. (2019). Wider face and pedestrian challenge 2018: Methods and results, *arXiv preprint arXiv: 1902.06854*.

Lu, Jiajun et al. (2017). Adversarial examples that fool detectors. *arXiv preprint arXiv: 1712.02494*.

Ma, B. et al. (2016). Visual tracking under motion blur. *IEEE Transactions on Image Processing*, 25(12), pp. 5867–5876.

Manjula, S. et al. (2016). A Study on Object Detection. *International Journal of Pharmacy & Technology*, (8), 22875–22885.

Marra, F. et al. (2018). On the vulnerability of deep learning to adversarial attacks for camera model identification. *Signal Processing: Image Communication*, 65, pp. 240–248.

Martins, N. et al. (2020). Adversarial machine learning applied to intrusion and malware scenarios: A systematic review. *IEEE Access 8*, pp. 35403–35419.

Massoli, F. V. et al. (2021). Detection of face recognition adversarial attacks. *Computer Vision and Image Understanding,* 202, 103103.

Menon, S., & Chapman, D. (2022). Semi-supervised Contrastive Outlier Removal for Pseudo Expectation Maximisation (SCOPE). *arXiv preprint arXiv: 2206.14261*.

Mijwil, M. M. et al. (2023). MobileNetV1-Based Deep Learning Model for Accurate Brain Tumour Classification. *Mesopotamian Journal of Computer Science*, 2023, pp. 32–41.

Moosavi-Dezfooli, Seyed-Mohsen et al. (2016). Deepfool: a simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582.

Moosavi-Dezfooli, Seyed-Mohsen et al. (2017). Universal adversarial perturbations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773.

Mukti, I. Z., & Dipayan, B. (2019). Transfer learning-based plant disease detection using ResNet50. *4th IEEE International Conference on Electrical Information and Communication technology (EICT),* pp. 1–6.

Mungoli, N. (2023). Scalable, Distributed AI Frameworks: Leveraging Cloud Computing for Enhanced Deep Learning Performance and Efficiency. *arXiv preprint arXiv: 2304.13738*.

Murthy, Chinthakindi Balaram et al. (2020). Investigations of object detection in images/videos using various deep learning techniques and embedded platforms - a comprehensive review. *Applied Sciences*, 10(9), 3280.

Nawaratne, Rashmika et al. (2020). Spatiotemporal Anomaly Detection Using Deep Learning for Real-Time Video Surveillance. *IEEE Transaction Actions on Industrial Informatics, 16 (1)*, pp. 393–402.

Nguyen, K., Fernando, T., Fookes, C., & Sridharan, S. (2023). Physical adversarial attacks for surveillance: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 35(12)*, 17036–17056.

Niculae, S. (2018). Reinforcement learning vs. genetic algorithms in game-theoretic cyber-security. *Faculty of Mathematics and Computer Science*.

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C. C., Lee, J. T., ... & Desai, M. (2011, June). A large-scale benchmark dataset for event recognition in surveillance video. *CVPR 2011*, pp. 3153–3160.

Ohira, Shuji et al. (2020). Normal and malicious sliding windows similarity analysis method for fast and accurate IDS against DoS attacks on in-vehicle networks. *IEEE Access,* 8, pp. 42422–42435.

OpenAI (2017). Attacking machine learning with adversarial examples. https://openai.com/research/attacking-machine-learning-with-adversarial-examples.

Ou, Weihua et al. (2014). Robust face recognition via occlusion dictionary learning. *Pattern Recognition*, 47(4), pp. 1559–1572.

Ozbulak, Utku et al. (2020). Perturbation analysis of gradient-based adversarial attacks. *Pattern Recognition Letters*, 135, pp. 313–320.

Pan, Guangyuan et al. (2019a). Evaluation of alternative pre-trained convolutional neural networks for winter road surface condition monitoring. *IEEE 5th International Conference on Transportation Information and Safety (ICTIS)*, pp. 614–620.

Pan, Zhaoqing et al. (2019b). Recent progress on generative adversarial networks (GANs): A survey. *IEEE Access*, 7, pp. 36322–36333.

Papernot, Nicolas et al. (2016). The limitations of deep learning in adversarial settings. *IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387.

Papernot, Nicolas et al. (2017). Practical black-box attacks against machine learning. *Proceedings of 2017 ACM on Asia Conference on Computer and Communications Security*, pp. 506–519.

Park, Yesul et al. (2021). Multiple object tracking in deep learning approaches: A survey. *Electronics,* 10(19), 2406.

Peng, Chunlei et al. (2020a). Soft semantic representation for cross-domain face recognition. *IEEE Transactions on Information Forensics and Security*, 16, pp. 346–360.

Peng, Jinjia et al. (2020b). Cross-domain knowledge learning with dual-branch adversarial network for vehicle re-identification. *Neurocomputing*, 401, pp. 133–144.

Phuc, Le Tran Huu et al. (2019). Applying the Haar-cascade algorithm for detecting safety equipment in safety management systems for multiple working environments. *Electronics,* 8(10), 1079.

Picek, S., & Domagoj, J. (2022). Evolutionary computation and machine learning in security. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 1572–1601.

Poursaeed, Omid et al. (2018). Generative adversarial perturbations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4422–4431.

Prakash, C. D., & Lina J. K. (2021). It GAN DO better: GAN-based detection of objects on images with varying quality. *IEEE Transactions on Image Processing*, 30, pp. 9220–9230.

Qiu, Shilin et al. (2019). Review of artificial intelligence adversarial attack and defence technologies. *Applied Sciences*, 9(5), pp. 1–29.

Radford, Alec et al. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv: 1511.06434*.

Raja, L., & Periasamy, P. S. (2022). A trusted distributed routing scheme for wireless sensor networks using block chain and jelly fish search optimizer based on deep generative adversarial neural network (Deep-GAN) technique. *Wireless Personal Communications*, 126(2), pp. 1101–1128.

Rajjak, S. S. A., & Kureshi, A. K. (2019). Recent advances in object detection and tracking for high resolution video: Overview and state-of-the-art. *IEEE 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, pp. 1–9.

Raschka, Sebastian et al. (2020). Machine learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, 11(4), 193.

Raskar, P. S., & Sanjeevani, K. S. (2021). Real-time object-based video forgery detection using YOLO (V2). *Forensic Science International*, 327, 110979.

Rawal, B. S., & Gunasekaran M. (2021). Implementation of a secure multi-cloud storage framework with next-generation cryptosystems and split-protocol. *2021 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6.

Redmon, Joseph et al. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788.

Ren, Kui et al. (2020). Adversarial attacks and defences in deep learning. *Engineering*, 6(3), pp. 346–360.

Ren, Shaoqing et al. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, https://arxiv.org/abs/1506.01497.

Robert, V. N. J., & Vidya, K. (2022). OAM-GANN: Online Adaptive Memory-Based Genetically Optimized Artificial Neural Network for PUEA Detection in CRN Applications. *Research Square*, pp. 1–22.

Roheda, S., Riggan, B. S., Krim, H., & Dai, L. (2018). Cross-modality distillation: A case for conditional generative adversarial networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2926–2930.

Roheda, Siddharth et al. (2020). Robust multi-modal sensor fusion: An adversarial approach. *IEEE Sensors Journal*, 21(2), pp. 1885–1896.

Rosenberg, Ishai et al. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5), pp. 1–36.

Ruiqiang, Z., Yu, Z., & Xin, J. (2021). Optimization of small-object detection based on generative adversarial networks. *E3S Web of Conferences*, *EDP Sciences, Vol. 245,* p. 03062.

Ryu, S. E., & Chung, K. Y. (2021). Detection model of occluded object based on YOLO using hard-example mining and augmentation policy optimization. *Applied Sciences*, 11(15), p. 7093.

Sahithi, Ambati et al. (2023). Enhancing Object Detection and Tracking from Surveillance Video Camera Using YOLOv8. *IEEE International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, pp. 228–233.

Salimans, Tim et al. (2016). Improved techniques for training GANs. *30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain*, pp. 1–9.

Sanchez-Patiño, N. et al. (2021). Convolutional Neural Networks for Chagas' Parasite Detection in Histopathological Images. *43rd Annual International*

Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 2732–2735.

Segalis, E. & Galili, E. (2020). OGAN: Disrupting deepfakes with an adversarial attack that survives training. *arXiv preprint arXiv: 2006.12247*.

Senthil Murugan, A. et al. (2018). A study on various methods used for video summarization and moving object detection for video surveillance applications. *Multimedia Tools and Applications*, 77(18), pp. 23273–23290.

Sharma, I. (2021). A More Responsible Digital Surveillance Future. *Federation of American Scientists, A Special Project on Emerging Technologies and International Security of the Federation of American Scientists*, 5.

Shetty, Anirudha B. et al. (2021). Facial recognition using Haar cascade and LBP classifiers. *Global Transitions Proceedings,* 2(2), pp. 330–335.

Shin, Jungsup et al. (2020). Fast and robust object tracking using tracking failure detection in kernelized correlation filter. *Applied Sciences*, 10(2), 713, pp. 1–13.

Siegel, Jonathan (2021). The importance of loss functions: A note on the evolution of the toxicity probability interval design. *Contemporary Clinical Trials Communications*, 22, 100694, pp. 1–3.

Simao, Miguel et al. (2019). Improving novelty detection with generative adversarial networks on hand gesture data. *Neurocomputing*, 358, pp. 437–445.

Simonyan, K. & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems,* 27.

Smeulders, Arnold W. M. et al. (2013). Visual tracking: An experimental survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), pp. 1442–1468.

Sohl-Dickstein, Jascha et al. (2015). Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning. Ed. by Francis Bach & David Blei.*, Vol. 37, pp. 2256–2265.

Song, Dawn et al. (2018). Physical adversarial examples for object detectors. *12th USENIX Workshop on Offensive Technologies (WOOT 18)*.

Song, Shaojian et al. (2021). A new real-time detection and tracking method in videos for small-target traffic signs. *Applied Sciences*, 11(7), 3061.

Sreeja, G. G. et al. (2023). Traffic Infraction and Alert System for Two-wheelers Using Deep Learning and YOLO v3. *IEEE 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1, pp. 1000–1005.

Su, Nan et al. (2022). Detect Larger at Once: Large-Area Remote-Sensing Image Arbitrary-Oriented Ship Detection. *IEEE Geoscience and Remote Sensing Letters 19*, pp. 1–5.

Sun, L., Dou, Y., Yang, C., Zhang, K., Wang, J., Yu, P. S., ... & Li, B. (2022). Adversarial attack and defense on graph data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *35*(8), 7693–7711.

Sun, Rui et al. (2019). Robust visual tracking based on a convolutional neural network with extreme learning machine. *Multimedia Tools and Applications*, 78, pp. 7543–7562.

Suthishni, D., Nethra, P., & Senthil Kumar K. S. (2022). A review on machine learning-based security approaches in intrusion detection system. *IEEE 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 341–348.

Szegedy, Christian et al. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv: 1312.6199*.

Taher, Fatma et al. (2022). A Novel Tunicate Swarm Algorithm with Hybrid Deep Learning Enabled Attack Detection for Secure IoT Environment. *IEEE Access*, 10, pp. 127192–127204.

TensorFlow (2020). Adversarial example using FGSM. https://www.tensorflow.org/tutorials/generative/adversarial_fgsm.

Thakur, N., & Li, B. (2022a). PAT: Pseudo-Adversarial Training for Detecting Adversarial Videos. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Los Alamitos, CA, USA: IEEE Computer Society*, pp. 130–137.

Thakur, N., & Li, B. (2022b). PAT: Pseudo-Adversarial Training for Detecting Adversarial Videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 131–138.

Thys, S. et al. (2019a). Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Los Alamitos, CA, USA: IEEE Computer Society*, pp. 49–55.

Thys, Simen et al. (2019b). Fooling automated surveillance cameras: adversarial patches to attack person detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Ting, Liu et al. (2021). Ship Detection Algorithm Based on Improved YOLO V5. *2021 6th International Conference on Automation, Control and Robotics Engineering (CACRE)*, pp. 483–487.

Truong, Thanh Cong et al. (2020a). Artificial intelligence in the cyber domain: Offense and defense. *Symmetry,* 12(3), 410.

Truong, T. C., Diep, Q. B., & Zelinka, I. (2020). Artificial intelligence in the cyber domain: Offense and defense. *Symmetry*, *12*(3), 410.

UAE-P, The United Arab Emirates Government portal (2022). The UN E-government Survey: The UAE's competitiveness.

Ulfa, D. K., & Dwi, H. W. (2017). Implementation of Haar cascade classifier for motorcycle detection. *IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pp. 39–44.

Ullah, Zaib et al. (2020). Applications of artificial intelligence and machine learning in smart cities. *Computer Communications*, 154, pp. 313–323.

Van Noord, N., & Postma, E. (2017). Learning scale-variant and scale-invariant features for deep image classification. *Pattern Recognition, 61*, pp. 583–592.

Velastin, Sergio A. et al. (2020). Detecting, tracking and counting people getting on/off a metropolitan train using a standard video camera. *Sensors*, 20(21), 6251.

Vorobeychik, Y., & Kantarcioglu, M. (2018). *Adversarial machine learning*. Morgan & Claypool Publishers.

Vorobjov, Denis et al. (2018). An effective object detection algorithm for high-resolution video by using a convolutional neural network. *15th International Symposium on Neural Networks, Minsk, Belarus, Springer*, pp. 503–510.

Vähäkainu, P., & Lehto, M. (2019). Artificial intelligence in the cyber security environment. *The 14th International Conference on Cyber Warfare and Security ICCWS2019*.

Wang, B. et al. (2020a). New algorithm to generate the adversarial example of an image. *Optik*, 207, 164477.

Wang, Junhua et al. (2021). YOLOv5_CSL_F: YOLOv5's Loss Improvement and Attention Mechanism Application for Remote Sensing Image Object Detection.

2021 International Conference on Wireless Communications and Smart Grid (ICWCSG), pp. 197–203.

Wang, Qiong et al. (2020b). Overview of deep-learning based methods for salient object detection in videos. *Pattern Recognition,* 104, p. 107340.

Wang, Yaxing et al. (2018). Transferring GANs: Generating images from limited data. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 218–234.

Wang, Zhengwei et al. (2020c). Synthetic-Neuroscore: Using a neuro-AI interface for evaluating generative adversarial networks. *Neurocomputing*, 405, pp. 26–36.

Wanjale, K. H. et al. (2013). Use of Haar Cascade Classifier for Face Tracking System in Real-Time Video. *International Journal of Engineering Research and Technology*, 2(4).

Wei, Xingxing et al. (2018). Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv: 1811.12641*.

Wei, Xingxing et al. (2019). Transferable adversarial attacks for image and video object detection. *The 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*, 954–960.

Wen, Yandong et al. (2016). Structured occlusion coding for robust face recognition. *Neurocomputing*, 178, pp. 11–24.

Cao, W., Yuan, J., He, Z., Zhang, Z., & He, Z. (2018). Fast deep neural networks with knowledge guided training and predicted regions of interests for real-time video object detection. *IEEE Access*, *6*, 8990–8999.

Wu, Dongya et al. (2019). Generative Adversarial Networks for Exo-Atmospheric Infrared Object Discrimination. *IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pp. 1–5.

Wu, Tian-Hao et al. (2021). Real-Time Vehicle and Distance Detection Based on Improved YOLO v5 Network. *3rd World Symposium on Artificial Intelligence (WSAI)*, pp. 24–28.

Xiao, Chaowei et al. (2018a). Generating adversarial examples with adversarial networks. *arXiv preprint arXiv: 1801.02610*.

Xiao, Chaowei et al. (2018b). Spatially transformed adversarial examples, *In:*

arXiv preprint arXiv: 1801.02612.

Xu, Bo et al. (2020). Adversarial attacks for object detection. *39th IEEE Chinese Control Conference (CCC)*, pp. 7281–7287.

Yadav, S., & Shahram, P. (2018). Understanding tracking methodology of kernelized correlation filter. *9th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 1330–1336.

Yang, Darren Yu et al. (2018). Building Towards Invisible Cloak: Robust Physical Adversarial Attack on YOLO Object Detector. *9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 368–374.

Yang, Hongju et al. (2019a). ContourGAN: Image contour detection with generative adversarial network. *Knowledge-Based Systems*, (164), pp. 21–28.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). Wider face: A face detection benchmark. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5525–5533.

Yang, Yumin et al. (2021). Remote Sensing Image Aircraft Target Detection Based on GIoU-YOLO v3. *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pp. 474–478.

Yang, Yuxing et al. (2019b). Enhanced adversarial learning-based video anomaly detection with object confidence and position. *13th IEEE International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–5.

Yazdinejad, A., Dehghantanha, A., Parizi, R. M., Hammoudeh, M., Karimipour, H., & Srivastava, G. (2022). Block Hunter: Federated learning for Cyber Threat Hunting in Blockchain-Based IoT Networks. *IEEE Transactions on Industrial Informatics*, *18*(11), 8356–8366.

Ye, Dong Hye et al. (2018). Deep learning for moving object detection and tracking from a single camera in unmanned aerial vehicles (UAVs). *Electronic Imaging*, 2018(10), pp. 466–1.

Yildirim, G., & Süsstrunk, S. (2015). FASA: fast, accurate, and size-aware salient object detection. *12th Asian Conference on Computer Vision, Singapore, Singapore, Springer*, pp. 514–528.

Yu, J., & Choi, H. (2021). YOLO MDE: Object detection with monocular depth estimation. *Electronics,* 11(1), 76.

Yu, Tingzhao et al. (2018). Deep generative video prediction. *Pattern Recognition Letters*, 110, pp. 58–65.

Yuan, Yue et al. (2020). A scale-adaptive object-tracking algorithm with occlusion detection. *EURASIP Journal on Image and Video Processing*, pp. 1–15.

Zahisham, Zharfan et al. (2020). Food recognition with ResNet-50. *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*, pp. 1–5.

Zancato, Luca et al. (2022). On the trainability and generalization of deep neural networks. *PhD Thesis, Department of Information Engineering University of Padova*.

Zawya.com (2016). UAE ranks 8th globally and 1st regionally in UNS 2016 e-smart services index. URL: https://www.zawya.com/en/business/uae-ranks-8th-globally-and-1st-regionally-in-uns-2016-e-smart-services- index-it2f4gip.

Zeng, Dan et al. (2021). A survey of face recognition techniques under occlusion. *IET Biometrics*, 10(6), pp. 581–606.

Zhang, Chengyuan et al. (2019a). Crossing generative adversarial networks for cross-view person re-identification. *Neurocomputing*, 340, pp. 259–269.

Zhang, Gaowei et al. (2020a). Cross-scale generative adversarial network for crowd density estimation from images. *Engineering Applications of Artificial Intelligence*, 94, 103777.

Zhang, Kejun et al. (2019b). No one can escape: A general approach to detect tampered and generated images. *IEEE Access*, 7, pp. 129494–129503.

Zhang, Shihao et al. (2017). Kill two birds with one stone: Boosting both object detection accuracy and speed with adaptive patch-of-interest composition. *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 447–452.

Zhang, X., & Luo, X. (2020). Exploiting defenses against GAN-based feature inference attacks in federated learning. *arXiv preprint arXiv: 2004.12571*.

Zhang, XiaoQi et al. (2022). GAN-based Abnormal Transaction Detection in Bitcoin. *IEEE 7th International Conference on Smart Cloud (Smart-Cloud)*, pp. 157–162.

Zhang, Yongqiang et al. (2019c). Detecting small faces in the wild based on generative adversarial network and contextual information, *Pattern Recognition*, 94, pp. 74–86.

Zhang, Yongqiang et al. (2020b). Multi-task generative adversarial network for detecting small objects in the wild. *International Journal of Computer Vision*, 128, pp. 1810–1828.

Zhao, Fan et al. (2021). A KCF-Based Incremental Target Tracking Method with Constant Update Speed. *IEEE Access*, 9, pp. 73544–73560.

Zhao, Z. Q. et al. (2019). Object detection with deep learning: A review. *arXiv 2018, arXiv preprint arXiv: 1807.05511*.

Zhen, Xinxin et al. (2020). A visual object tracking algorithm based on improved TLD, *Algorithms*,13(1), 15.

Zheng, Hai-Tao et al. (2018). Automatic Generation of News Comments Based on Gated Attention Neural Networks, *IEEE Access*, 6, pp. 702–710.

Zheng, Yu et al. (2020a). An effective adversarial attack on person re-identification in video surveillance via dispersion reduction. *IEEE Access*, 8, pp. 183891–183902.

Zhihuan, Wu et al. (Apr. 2018). Rapid target detection in high-resolution remote sensing images using the YOLO model, *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-3*, pp. 1915–1920.

Zhong, Junping et al. (2020). Adversarial reconstruction based on tighter oriented localisation for catenary insulator defect detection in high-speed railways. *IEEE Transactions on Intelligent Transportation Systems*, 23(2), pp. 1109–1120.

Zhu, Aiyu et al. (2023). Deep Reinforcement Learning for Real-Time Assembly Planning in Robot-Based Prefabricated Construction. *IEEE Transactions on Automation Science and Engineering* 20(3), pp. 1515–1526.

Zhu, Haidi et al. (2020). Moving object detection with deep CNNs, *IEEE Access 8*, pp. 29729–29741.

Zhu, Jun-Yan et al. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232.

Zhu, Yukun et al. (2015). Segdeepm: Exploiting segmentation and context in deep neural networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4703–4711.