National College of Ireland

Project Submission Sheet

Student Name:	YOGESH SAVULLA		
Student ID:	23322004		
Programme:	Master Of Science In Data Analytics	**Year:**	2024 - 2025
Module:	Research Practicum		
Lecturer:	David Hamill		
Submission Due Date:	11/08/2025		
Project Title:	Deep Learning for Crowd Behavioral Analysis and Intelligent Management		
Word Count:	6342		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: S Yogesh

Date: 11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**

5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

[Insert Module Name]

[Insert Title of your assignment]

Your Name/Student Number	Course	Date

AI Acknowledgment

Tool Name	Brief Description	Link to tool

Description of AI Usage

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

Evidence of AI Usage

Additional Evidence:

Additional Evidence:

National College of Ireland

DECLARATION OF ETHICS CONSIDERATION

School of Computing

Student Name: SAVULLA YOGESH
Student ID: 23322004

Programme Msc in Data Analytics **Year:** 2024-2025

Module: Research Practicum

Project Title: Deep Learning for Crowd Behavioral Analysis and Intelligent Management

Please circle (or highlight) as appropriate

This project involves human participant	**Yes** / No

Introduction

Secondary data refers to data that is collected by someone other than the current researcher. Common sources of secondary data for social science include censuses, information collected by government departments, organizational records and data originally collected for other research purposes. **Primary data**, by contrast, is collected by the investigator conducting the research.

Target system refers to any asset, device, infrastructure component, hosting environment, network, or application, that is intended to be used by the current researcher as the target system for testing their ICT solution artefact. This can include a penetration testing target or any system under test.

Rules of Engagement is a document that contains the formal approvals, authorizations, scope, and other general guidelines or formal objectives necessary to execute a penetration testing engagement.

The following decision table will assist you in deciding if you have to complete the Declaration of Ethics Consideration Form or/and the Ethics Application Form.

Public Data/Target System	Y	Y	Y	Y	N	N	N	N
Private Data/Target System	Y	Y	N	N	Y	Y	N	N
Human Participants	Y	N	Y	N	Y	N	Y	N
Declaration of Ethics Consideration Form	X	X	X	X	X	X	X	
Ethics Application Form	X		X		X		X	

A project that does not involve human participants requires ONLY completion of Declaration of Ethics Consideration Form and submission of the form on the module's Moodle page.

A project that involves human participants requires ethical clearance. Thus, it requires the completion of both a Declaration of Ethics Consideration Form AND an Ethics Application Form. These must be submitted through the module's Moodle page. Please refer to and ensure compliance with the ethical principles stated in NCI Ethics Form available on the Moodle page.

Please circle (or highlight) as appropriate.

The project makes use of secondary dataset(s) or target system(s) created by the researcher	**Yes** / No
The project makes use of public secondary dataset(s) or target system(s)	**Yes** / No
The project makes use of non-public secondary dataset(s) or target system(s) Approval letter from non-public secondary dataset(s) or target system(s) owner received	Yes / **No** Yes / No

Sources of Data/Target System

It is students' responsibility to ensure that they have the correct permissions/authorizations to use any data or target system in a study. Projects that make use of data or target system that does not have authorization to be used, will not be graded for that portion of the study that makes use of such data or target system. For target systems, the student must ensure that the asset is used in accordance with the company's acceptable usage policies and Terms of Use.

Public Data / Public Target System

A project that makes use of public secondary dataset(s) or public target system(s) ***does not need ethics permission*** *but* ***needs a letter/email from the copyright holder/ target system owner*** *regarding potential use.*

Some websites and data sources allow their data sets or platforms to be used under certain conditions. In these cases, a letter/email from the copyright holder is NOT necessary, but the researcher should cite the source of this permission and indicate under what conditions the data or target system are allowed to be used. See Appendix I for examples of permissions for data granted by Fingal Open Data, and Eurostat website. See Appendix III for examples of permissions granted by some penetration testing target systems.

Where websites or data sources indicate that they do not grant permission for data or their platform to be used, you will still need a letter/email from the copyright holder. For example, see Appendix II for an example from the Journal of Statistics Education.

Private Data / Private Target System (not created by the researcher)

A project that makes use of non-public (private) secondary dataset(s) or target system must receive usage permission from the School of Computing.

An approval letter/email from the owner *(e.g. institution, company, etc.)* ***of the non-public secondary dataset or target system must be attached to the Declaration of Ethics Consideration****. For datasets, the letter/email must confirm that the dataset is anonymised and permission for data processing, analysis and public dissemination is granted.*

For target systems, the letter/email must provide evidence that permission/authorization has been granted to the student for using the system. In addition to this, in the cases where the goal of the project is to conduct penetration testing, vulnerability scans, or other ethical hacking engagement, the student must also provide a document describing the rules of engagement.

Evidence to include for use of secondary dataset(s) / target system(s)
Include dataset(s) / target system(s) owner letter/email or cite the source for usage permission.

The Dataset which I used its tentative and still yet not decided at this point.

CHECKLIST

Non-public/private secondary dataset(s)/target system(s) - Dataset / target system Owner letter/email - Rules of Engagement if applicable see Appendix IV - Rules of Engagement Worksheet) is attached to this form ***OR*** Citation and link to the website where permission is granted – provided in this form	 Yes / No Yes / No **Yes** / No

ETHICS CLEARANCE GUIDELINES WHEN HUMAN PARTICIPANTS ARE INVOLVED

The Ethics Application Form must be submitted on Moodle for approval prior to conducting the work.

Considerations in data collection

- Participants will not be identified, directly or through identifiers linked to the subjects in any reports produced by the study
- Responses will not place the participants at risk of professional liability or be damaging to the participants' financial standing, employability or reputation
- No confidential data will be used for personal advantage or that of a third party

Informed consent

- Consent to participate in the study has been given freely by the participants
- participants have the capacity to understand the project goals.
- Participants have been given information sheets that are understandable

- Likely benefits of the project itself have been explained to potential participants
- Risks and benefits of the project have been explained to potential participants
- Participants have been assured they will not suffer physical stress or discomfort or psychological or mental stress
- The participant has been assured s/he may withdraw at any time from the study without loss of benefit or penalty
- Special care has been taken where participants are unable to consent for themselves (e.g children under the age of 18, elders with age 85+, people with intellectual or learning disability, individuals or groups receiving help through the voluntary sector, those in a subordinate position to the researcher, groups who do not understand the consent and research process)
- Participants have been informed of potential conflict of interest issues
- The onus is on the researcher to inform participants if deception methods have to be used in a line of research

I have read, understood, and will adhere to the ethical principles described above in the conduct of the project work.

Signature: S Yogesh

Date: 11/04/2025

Appendix I

1) Fingal Open Data: http://data.fingal.ie/About

Licence
Citizens are free to access and use this data as they wish, free of charge, in accordance with the Creative Commons Attribution 4.0 International License (CC-BY).

Note: From November 2010 to July 2015, data on Fingal Open Data was published in accordance with the PSI general licence.

Use of any published data is subject to Data Protection legislation.

Licence Statement
Under the CC-BY Licence, users must acknowledge the source of the Information in their product or application by including or linking to this attribution statement: "Contains Fingal County Council Data licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence".

Multiple Attributions
If using data from several Information Providers and listing multiple attributions is not practical in a product or application, users may include a URI or hyperlink to a resource that contains the required attribution statements.

2) Eurostat: https://ec.europa.eu/eurostat/about/policies/copyright

COPYRIGHT NOTICE AND FREE RE-USE OF DATA

Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed below, can be reused without any payment or written licence provided that:

- the source is indicated as Eurostat
- when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information

Appendix II

Journal of Statistics Education: http://jse.amstat.org/jse_users.htm

JSE Copyright and Usage Policy
Unlike other American Statistical Association journals, the Journal of Statistics Education (JSE) does not require authors to transfer copyright for the published material to JSE. Authors maintain copyright of published material. Because copyright is not transferred from the author, permission to use materials published by JSE remains with the author. Therefore, to use published material from a JSE article the requesting person must get approval from the author.

National College of Ireland

Deep Learning for Crowd Behavioural Analysis and Intelligent Management

MSc Research Project

MSc Data Analytics

Yogesh Savulla

Student ID: 23322004

School of Computing

National College of Ireland

Supervisor: Prof. David Hamill

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	YOGESH SAVULLA .....
Student ID:	23322004 .....
Programme:	MSc Data Analytics ..... **Year:** .....2025...
Module:	MSc Research Project .....
Supervisor:	David Hamil .....
Submission Due Date:	11/08/2025 .....
Project Title:	Deep Learning for Crowd Behavioral Analysis and Intelligent Management
Word Count:	.....6342..... **Page Count**.....23.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Yogesh savulla .....
Date:	11 /08/2025 .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deep Learning for Crowd Behavioural Analysis and Intelligent Management

Yogesh Savulla

23322004

Abstract

Anomaly detection in crowds is a key domain of research in the field of computer vision with applications to crowd safety, event surveillance and smart transportation. The conventional surveillance is based upon the manual observation that is inaccurate and not efficient in the implementation at large scale. Increased research into deep learning, specifically convolutional and attention-based networks, have made it possible to model video signals in terms of spatiotemporal dependencies automatically and make those models more robust in status as odd behaviour among the crowd.

Design A hybrid Convolution + Transformer-based video autoencoder that learns through a dataset of video sequences an understanding of normal crowd behaviour and can identify abnormality by reconstructing the input and analysing the perceptual error. The model can capture such short-term temporal dependencies by using sequences of five frames inPed1 and Ped2 UCSD Pedestrian Datasets to predict the following frame.

Mean Squared Error (MSE), Mean Absolute Error (MAE), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR) are used to measure reconstruction quality. They are fed into LightGBM as classifier features to dictate anomaly occurrence. It had a high reconstruction fidelity (MSE: 0.000282, SSIM: 0.9821, PSNR: 36.55 dB), and robust classification accuracy (accuracy: 77.84%, precision: 0.5845, recall: 0.6474, F1-score: 0.6144).

The results show how using deep feature extraction on the temporal dynamics combined with a learned decision boundary on the perceptual metrics to achieve strong annotation-saving performance-based anomaly detection. This method can find its way in real time application in detecting unattended objects, unlawful access, or strange movements. Advancement work in the future may involve the use of object tracking, adaptation of an objective within different camera environments, and multimodal sensor fusion to enhance detection capability further and flexibility.

1 Introduction

The pressing necessity of safer streets and more effective means of security has paved the way to crowd anomaly detection emerging as a serious area of computer vision. During this fast-growing urbanization as crowds of people gather in transport pickups, large and medium-sized stadiums, shopping malls and areas of cities, it becomes even riskier and more imperative that it should be duly observed and monitored. The elder surveillance approaches

using human operators cannot be expanded to cover large networks of cameras and have a high likelihood of operator fatigue leading to failure to monitor. Such issues have prompted embracement of smart surveillance technologies that use deep learnings to automate the process of spotting anomalies. As opposed to the previously employed traditional methods whereby the spatial or time patterns were usually studied independently of each other, contemporary practices aim to combine the two to enable a deeper spatiotemporal sense. The proposed robust hybrid Convolution + Transformer video autoencoder with multi-metric decision framework can recognise abnormal crowd behaviour under various conditions, which provides a scalable and computationally efficient counterpart to the video surveillance practice with high performance.

1.1 Background and Motivation

Intelligent surveillance system requires an urgent solution in regard to the detection of anomalous events within the crowd setting. As the trend of urbanization continues to grow, mass events take place in plain sight areas like transportation networks, sports arena, shopping centers, and urban streets. It is important to monitor these environments in order to detect any abnormalities that may cause accidents or threats to the people.

Conventional surveillance systems largely rely on the human operators to manually monitor camera footage. It is a type of method that requires a great number of resources, is easily susceptible to human error through fatigue and cannot easily scale up in hundreds or thousands of live video streams. These constraints have spurred the use of computer vision methods that will have the ability to automate the identification of abnormalities. (Dardour et al., 2025)

In the specific context of crowd behaviour, anomalies can manifest in various forms:

- Object-based anomalies, such as bicycles or carts in pedestrian-only zones
- Behavioural anomalies, like running against the flow of pedestrians
- Environmental anomalies, including sudden gatherings or dispersals

Detecting such anomalies is inherently challenging because abnormal events are rare, highly varied, and context-dependent. Furthermore, in many real-world settings, anomaly labels are scarce or unavailable, making supervised classification difficult to apply directly.

1.2 Research Gap

Deep learning, and convolutional neural networks (CNNs) and recurrent models in particular, have changed video analysis over the last decade. Nevertheless, these methods usually concentrate on solely spatial or temporal characteristics that restrict them in finding complicated spatiotemporal correlations. More recently Transformers have been promising at modelling long range dependent data based on a sequential input, however pure attention models can be computationally expensive and are, when used directly on high-resolution video data, computationally demanding (Zeghina et al., 2024).

The other gap in the current research is the robustness of decisions. A large number of anomaly detection systems are based on one metric, e.g., reconstruction error, using a constant threshold. The approach is robust to variations in scenes, shifts in lighting and benign deviations of the training distribution. More powerful would be a combination of

more than one perceptual quality measure and learning the decision boundary by example (Paolini et al., 2025).

1.3 Research Objectives

The objectives of this work are:

- To design a hybrid Convolution + Transformer Video Autoencoder capable of modeling both spatial features and short-term temporal dependencies for crowd behavior analysis.
- To evaluate the reconstruction quality of the autoencoder using both pixel-based and perceptual metrics (MSE, MAE, SSIM, PSNR).
- To integrate a meta-classifier (LightGBM) that uses multiple reconstruction metrics as features for binary anomaly classification.
- To validate the proposed system on benchmark datasets (UCSD Ped1 and Ped2) and compare performance against baseline thresholds.

1.4 Contribution to the Scientific Literature

This study contributes to the field of crowd anomaly detection in three key ways:

- Architectural innovation: We propose a lightweight Conv+Transformer encoder-decoder that balances accuracy and computational efficiency.
- Two-stage detection pipeline: By separating representation learning from anomaly decision-making, we enable more flexible calibration and improve robustness to scene changes.
- Comprehensive evaluation: We present quantitative and qualitative results from the UCSD datasets, demonstrating how perceptual metrics complement pixel error measures for anomaly detection.

1.5 Structure of the Report

The rest of this paper is organized as follows:

- Section 2: Related Work — Provides a critical review of existing anomaly detection methods, including reconstruction-based, predictive, and hybrid approaches.
- Section 3: Research Methodology — Describes the experimental setup, dataset preprocessing, model architecture, and evaluation protocol.
- Section 4: Design Specification — Details the design considerations, architectural choices, and system requirements for the proposed approach.
- Section 5: Implementation — Outlines the implementation details, including tools, frameworks, and code structure.
- Section 6: Evaluation — Presents the results, analysis, and discussion, supported by figures and tables.
- Section 7: Conclusion and Future Work — Summarizes the findings, discusses limitations, and outlines potential directions for future research.

2 Related Work

The problem of detecting crowd anomalies has a long history, starting with the handcrafted features extraction approaches and moving onto the deep learning architectures which can directly learn pixel-level features and can be trained end-to-end. The following

section entails a critical literature survey of three related areas, including reconstruction-based anomaly detection, predictive modeling, and hybrid architectures using convolutional and attention mechanisms. We also talk about decision-making strategies under reconstruction metrics and the combination of meta-classifiers in order to produce powerful anomaly identification (Sharif et al., 2025).

2.1 Reconstruction-Based Approaches

Learning a model, capable of faithfully reconstructing normal patterns of motion and appearance, is one of the most popular approaches to anomaly detection in video, and uses training data which does not contain anomalous data. The implicit hypothesis is that anomalous patterns will be poorily reconstructed by the model, which initially did not see such patterns during training, and so reconstruction errors are high and can be exploited as an anomaly indicator (Wang & Yang, 2022).

The early techniques employed sparsely coded and dictionary learning, where typical patterns were encoded as sparse coefficients of a small set of basis vectors trained on normal training sets. Such methods were sensitive to noise, and although interpretable, they were hard to apply to complex motion patterns.

Autoencoders are the most popular reconstruction-based technique to emerge with deep learning. Convolutional Autoencoders (CAEs) have been proposed to take advantage of video frame spatial locality. The encoder will project the input frames into a lower dimension latent representation and the decoder will take the representation and try to reconstruct the input. The anomaly score is reconstruction error which is usually calculated through Mean Squared Error (MSE) (Wu et al., 2021).

But pure spatial CAEs do not consider temporal dependencies that are important in finding abnormalities in dynamic scenes. To give an example, when one is walking, it is normal but then when the same person is running in a crowd, it may be anomalous. Temporal-unaware spatial models may not be able to identify such instances since the frozen look of the individual is not necessarily abnormal.

In order to add temporal context, scientists have augmented autoencoders by adding Recurrent Neural Networks (RNNs) or a Long Short-Term Memory (LSTM) layer, yielding ConvLSTM Autoencoders to model short-term dynamics. The models are able to deal with predictable temporal sequences but have difficulties in long-range dependencies because of the shortcomings of recurrent networks.

Another addition is predictive autoencoders, in which the model learns how to predict future frames based on the past frames, as opposed to simply reconstruction of the input. Predictive reconstruction compels the model to learn about motion and the evolution of time better, hence being more likely to detect anomalies that complicate the assumed pattern of motions (Yang & Radke, 2025).

In spite of their success, most reconstruction-based approaches require only a single anomaly score computed with respect to MSE or other metrics and are therefore vulnerable to benign lighting, camera angle or background clutter variations. This encouraged the research into the multi-metric and meta-classifiers, the employment of which has been carried out in the current study (Luo & Zhang, 2023).

2.2 Predictive and Hybrid Architectures

Anomaly detection by predictive modeling makes use of video sequence temporal properties. The model can predict the next frame instead of reconstructing an existing one on a series of prior frames. Such a configuration automatically represents both movement and temporal continuity: when the following picture is far off the forecast, there may be an anomaly.

Optical Flow was applied as an intermediate representation in early predictive models, in which the flow field between adjacent frames was predicted. This benefited in that it was not concerned with raw pixel values but instead motion patterns, however, this came at the cost of added complexity to computing and interpreting the flow vectors and especially dense crowd scenes (P & K, 2025).

In deep learning, 3D Convolutional Neural Networks (3D-CNNs) gained popularity as a means of extracting spatiotemporal characteristics in the same architecture. The 3D-CNNs have the advantage of directly modelling short sequences of frames by convolving in both spatial and temporal dimensions. Nevertheless, they are computationally expensive and memory-intensive and thus less likely to be used in the real-time or resource-limited scenarios.

Transformer architectures on video analysis was another line of research. Transformers that were originally used in Natural Language Processing are good at long-range dependencies through self-attention. When applied to video, they can record complicated motion patterns with long curves of time. But self-attention has a quadratic computational cost with input size and the length of its sequence, the cost of which is prohibitive in high-resolution videos (Dhar et al., 2025).

As an answer to this, hybrid Conv+Transformer architectures have been proposed. Convolution layers in these models downsample and decompositions the spatial representations of inputs into compact tokens that the Transformer can process. The combination maintains the efficiency of convolutions to spatial encoding and the ability of attention mechanisms to do temporal reasoning.

Hybrid architectures even have the potential to provide a trade-off between efficiency and accuracy in the case of anomaly detection. The convolutional front-end guarantees effective capture of the spatial characteristics which include the crowd density, the shape of the object and the layout of the scene. Transformer back-end learns how these features change over time and potential anomalies in behavior can be identified.

This is a hybrid of our work: a three-stage convolutional encoder containing residual refinement of the spatial information, and a lightweight Transformer block that processes temporal sequences of feature maps. The design is computationally cheap but is also highly capable in terms of temporal modeling (Xu et al., 2021).

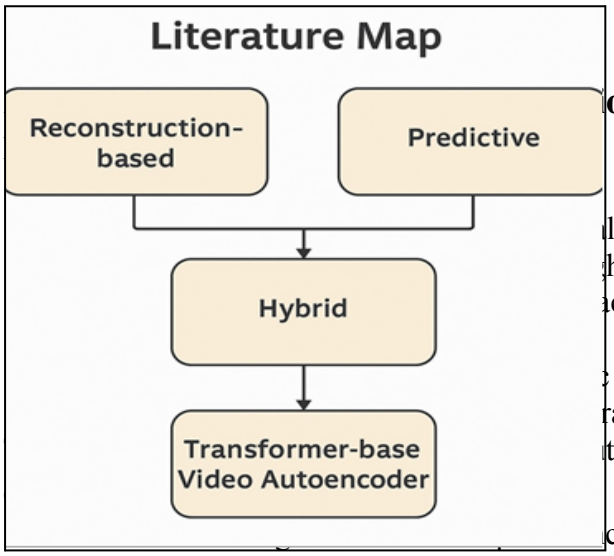
2.3 Metric-Based Decision Strategies

Most reconstruction or predictive models obtain an anomaly score as a direct mapping of a single error measure, e.g. MSE between the reconstructed/predicted frame and ground truth. This is simple to do, but brittle. One example is a sudden fluctuation in lighting resulting in a large MSE but not representing a true anomaly resulting in false positives.

Multi-metric fusion has been explored in order to enhance robustness. Other perceptual measures besides MSE include Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) that can be used to capture the varying aspects of reconstruction quality. SSIM is interested in structural transformations that are more humanized and PSNR is considered to be a global measure of fidelity (Leist et al., 2022).

Using these metrics together gives a more detailed account of reconstruction quality. Rather than having to preset thresholds on all the metrics, a machine learning classifier can be trained to find the best decision boundary in the multiple dimensional space of metrics. This method will be able to adjust to the properties of a particular dataset and accommodate change in normal patterns to a larger degree.

The philosophy is incorporated in our approach where we calculate MSE, MAE, 1-SSIM, and -PSNR of every reconstructed frame. The input to a LightGBM classifier is these four features and it is trained on the data in the UCSD test sets to learn the difference between normal and anomalous frames by using labeled ground truth masks. The decision boundaries that are learned by the classifier are more flexible than fixed-threshold approaches (Kenyeres et al., 2024).



on for Current Work

reconstruction-based methods lack strong
l models neglect spatial details.
h-capacity models like pure Transformers or
cy but are computationally prohibitive for

thresholds are sensitive to benign variations,
rade-offs.
toencoder with a LightGBM meta-classifier

coding with Transformer temporal modeling.

- Using a lightweight architecture to maintain computational efficiency.
- Employing a multi-metric decision stage to improve robustness to scene variations.

3 Research Methodology

The mentioned section describes the full lifespan of the research efforts that were conducted to design, train, and evaluate the proposed video anomaly detection framework. The initial stage will involve the UCSD Ped1 and Ped2 datasets that hold grayscale surveillance videos of the pedestrian walkways. Short-term temporal dynamics was captured by dividing raw footage into sequences of five successive frames of fix length. Preprocessing involved the frame resizing, normalization and changing the format to a tensor format that deep learning models could accept (Duong et al., 2023).

Its core model is a Convolutional Encoder-Decoder with an additional block of Transformer to model the temporal pattern. An encoder transforms spatial information on every frame, and the Transformer does sequential embeddings to reflect inter-frame connections. The decoder recreates the original frames based on the latent representation allowing to detect anomalies based on the quality of reconstruction.

Training was made using Adam optimizer and mean squared error loss, early stopping was used to inhibit overfitting. After reconstruction, various measures, namely, MSE, MAE, SSIM and PSNR, were calculated on sequence-by-sequence basis. Such characteristics were inserted into a LightGBM classifier to judge whether they were normal or caused anomalies (Ferrer, 2024).

The performance was assessed with accuracy, precision, recall, and F1-score and scientifically assessed through reconstruction error distributions. This end-to-end process enables spatial and temporal anomalies to be captured effectively as well as gives a quantitative foundation upon which to measure performance.

3.1 Research Questions and Hypotheses

The research is guided by two key questions:

- RQ1: Can a lightweight Conv+Transformer video autoencoder, trained exclusively on normal crowd behavior, produce reconstruction/perceptual error patterns that distinguish normal from anomalous events in UCSD crowd videos?
- RQ2: Can a meta-classifier train on multiple reconstruction/perceptual metrics outperform fixed-threshold anomaly detection based on a single metric?

From these, we hypothesize:

- H1: The hybrid Conv+Transformer architecture will yield low reconstruction error for normal frames and significantly higher error for anomalous frames.
- H2: A LightGBM classifier trained on multiple metrics (MSE, MAE, SSIM, PSNR) will improve overall detection accuracy and F1-score compared to thresholding any single metric.

3.2 Dataset

We use the UCSD Anomaly Detection Dataset v1p2, comprising Ped1 and Ped2 sequences:

- Ped1: 34 training and 36 testing videos captured at 158 × 238 resolution in a walkway scene with pedestrians moving in a constrained path.
- Ped2: 16 training and 12 testing videos captured at 240 × 360 resolution in a different walkway scene with similar crowd patterns (Orvile, 2025).

The typical walkers in a normal frame are pedestrians; the anomalous elements comprise of bicycles, carts, skaters, and vehicles that get to the pedestrian zones.

For uniformity and computational efficiency:

- All frames are resized to 128 × 128 pixels.
- Frames are converted to RGB to preserve color cues, even though UCSD videos are relatively low color-contrast.
- Sequence length T = 5: the model takes five consecutive frames as input and predicts the next frame.

The training set consists of only normal sequences from Ped1 and Ped2. The testing set consists of normal and anomalous sequences, pixel-level ground-truth masks indicating anomalies.

3.3 Data Preprocessing

The preprocessing pipeline:

1. Frame extraction from each video file into individual image files.
2. Resizing frames to 128 × 128 pixels using bilinear interpolation.
3. Sliding-window sequence generation: create overlapping sequences of length 5, with the target as the 6th frame.
4. Normalization: scale pixel values to the [0,1] range.
5. Mask alignment for test sequences: load binary masks (if present), resize to 128 × 128, and align with the target frame.

The data loaders are implemented in a custom UCSDDataset PyTorch class, handling sequence creation, mask association, and tensor formatting.

3.4 Model Architecture

The Video Autoencoder consists of:

Encoder:

- ConvTransformerBlock (3→32 channels, stride=2) + ResidualBlock
- ConvTransformerBlock (32→64, stride=2) + ResidualBlock
- ConvTransformerBlock (64→128, stride=2) + ResidualBlock

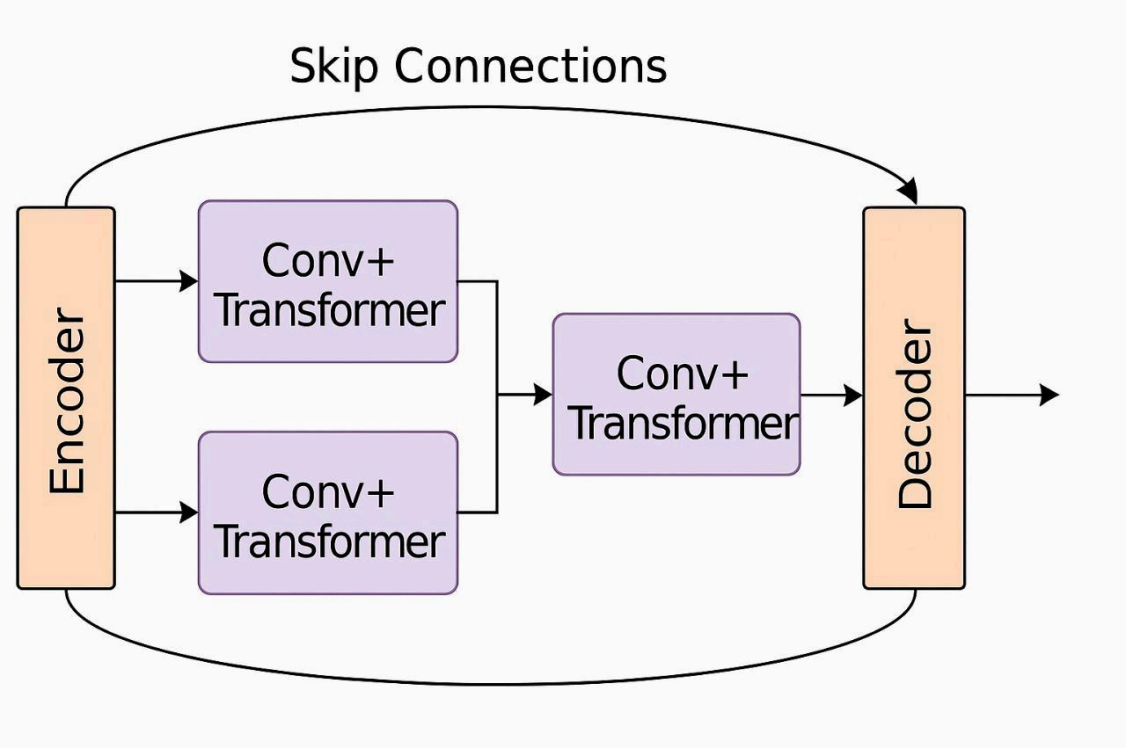
Each ConvTransformerBlock applies a stride-2 convolution for downsampling, a LayerNorm, and a Transformer encoder layer with 4 heads and a feedforward dimension 4× the channel size. The Transformer operates over temporal tokens formed from each spatial location across the T input frames (Olu-Ipinlaye & Mukherjee, 2024).

Decoder:

- ConvTranspose2d (128→64) + ResidualBlock
- ConvTranspose2d (64→32) + ResidualBlock
- ConvTranspose2d (32→3) with sigmoid activation

Skip connections from early encoder layers to corresponding decoder stages preserve spatial detail.

Output: Predicted next frame (\hat{x}_{t+1}) and an anomaly map computed as the mean squared residual per pixel.



3.5 Training Configuration

- Optimizer: Adam, learning rate 1×10^{-4}
- Loss function: Mean Squared Error (MSE) between predicted and ground-truth frames
- Batch size: 8
- Epochs: up to 30 with early stopping (patience = 5)
- Hardware: NVIDIA GPU (CUDA acceleration)

The training and validation sets are formed by an 80/20 split of the combined Ped1+Ped2 training sequences.

3.6 Evaluation Metrics

Four metrics quantify reconstruction quality:

- MSE: pixel-wise mean squared error
- MAE: pixel-wise mean absolute error
- SSIM: structural similarity index, values closer to 1 indicate higher perceptual similarity
- PSNR: peak signal-to-noise ratio in decibels, higher is better

For anomaly classification, we compute per-frame:

- Feature vector: [MSE, MAE, 1–SSIM, –PSNR]
- Train a LightGBM classifier with:
- n_estimators = 40
- scale_pos_weight = 2 (to handle anomaly/normal imbalance)
- 60/40 stratified train/test split

Classification performance is measured by:

- Accuracy

- Precision
- Recall
- F1-score
- Confusion matrix

3.7 Statistical Analysis

Train and test sets are also calculated in terms of mean and standard deviation of the metrics. Macro and weighted averages are reported as a classification figure. We discuss the precision-recall trade-off, and the contribution of individual metrics through LightGBM feature importance (Kashyap, 2024).

4 Design Specification

In this section, the design specification of the proposed anomaly detection system is provided and includes the description of its objectives, architecture, data flow, requirements and design rationale. It is designed to provide high accuracy, computational efficiency and robustness in detecting abnormalities in the crowd scene, with scalability in a wide variety of settings. At the heart of it is a hybrid Conv+Transformer architecture which uses convolutional layers to extract spatial information and Transformer layers to model time. The decision combines several reconstruction quality measurements into a LightGBM classifier, without using brittle fixed thresholds. The architecture is a compromise between performance and efficiency, and can run inference at 128x128 resolution in real time but can also support a future upgrade in architecture when sufficient resources are available (Cavallaro et al., 2023).

4.1 System Goals

The design goals for the proposed anomaly detection system were:

1. Accuracy: Detect a wide range of anomalies in crowd scenes.
2. Efficiency: Operate at low resolution (128×128) for faster inference.
3. Robustness: Avoid brittle fixed thresholds by learning a decision boundary.
4. Scalability: Maintain performance when trained on multiple scenes.

4.2 Architectural Choices

- Hybrid Conv+Transformer: Convolutions efficiently extract local spatial features and reduce dimensionality before the Transformer layer models temporal dependencies.
- Residual Refinement: Residual blocks after each ConvTransformer stage stabilize learning and preserve low-level details.
- Skip Connections: Improve reconstruction quality by reintroducing fine spatial features into the decoder.
- Multi-Metric Decision Stage: Using multiple reconstruction/perceptual metrics addresses the weakness of single-metric thresholding.

4.3 Data Flow and Processing

The runtime flow:

- Input: Sequence of 5 frames → preprocessed to 128×128, normalized.
- Encoder: Downsamples and encodes into a temporal latent representation.
- Transformer: Processes temporal tokens to model dynamics.
- Decoder: Reconstructs the next frame.
- Metrics: Compute MSE, MAE, SSIM, PSNR between reconstruction and target.
- LightGBM: Classifies frame as normal/anomalous based on metrics (Mnasri, 2019).

4.4 Requirements

- Software: PyTorch, Torchvision, OpenCV, PIQ (for SSIM/PSNR), LightGBM, scikit-learn
- Hardware: GPU with ≥8GB memory recommended for training
- Storage: ~10GB for datasets, models, and intermediate outputs

4.5 4.5 Justification of Design

Use of short sequence length (T=5) and 128x128 resolution is a tradeoff between temporal context and computation load. The multi-metric LightGBM phase enables the process of adjusting to the statistical characteristics of the data set without the manual tuning of the thresholds. Its architecture is modular: one can replace the encoderdecoder with one of a higher capacity in case of available resources.

Table 1: Summary of Model Parameters

Parameter	*Value*
Input size	$5 \times 3 \times 128 \times 128$
Encoder channels	32, 64, 128
Transformer heads	4
Batch size	8
Learning rate	1×10^{-4}
Epochs	≤ 30
Loss function	MSE
Classifier	LightGBM (n_estimators=40, scale_pos_weight=2)

5 Implementation

5.1 Development Environment

The whole pipeline was enabled in Python with PyTorch as a framework to elegantly perform deep learning, Torchvision to process and transform the datasets, open CV in order to manipulate the frames, as well as PIQ to measure the perceptual image quality (SSIM, PSNR). The meta-classification section applied LightGBM through the official Python API, and scikit-learn was applied to calculate the metrics, scale features and divide data.

The experiments were conducted in a GPU-enabled environment with CUDA acceleration. The key package versions included:

- PyTorch 2.6.0 + cu124
- Torchvision 0.21.0
- OpenCV 4.11.0
- PIQ 0.8.0
- LightGBM 4.6.0
- scikit-learn 1.2.2

5.2 Code Structure

The implementation was modularized into the following Python files:

- dataset.py: Implements the UCSDDataset class for loading sequences and optional masks.
- model.py: Contains the ConvTransformerBlock, ResidualBlock, and VideoAutoencoder definitions.
- train.py: Handles model training, validation, and early stopping logic.
- metrics.py: Computes MSE, MAE, SSIM, and PSNR for reconstruction evaluation.
- classifier.py: Trains and evaluates the LightGBM meta-classifier.
- utils.py: Helper functions for plotting, saving models, and logging metrics.

5.3 Training Procedure

We started the training procedure by loading the union Ped1 and Ped2 training sets. torch.utils.data.random_split was used to divide the unified dataset into subsets with an 80/20 ratio of the training and validations subsets.

The maximum number of epochs used during training was 30, and the early stopping mechanism was introduced in case the validation loss was not decreasing during 5 consecutive epochs. The Adam optimizer was set to have a learning rate of 1e-4 and reconstruction quality was optimized using MSE loss.

Since 5 frames were used as input sequences with a resolution of 128 x 128, then the batch size was allocated to 8 to fit the GPU memory limitation. The training loop repeated itself by switching forward passes, loss calculation, backpropagation and optimizer updates.

5.4 Model Saving and Loading

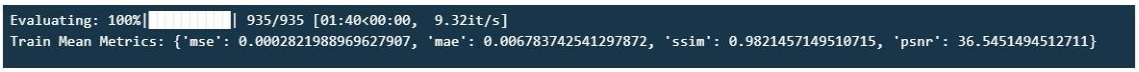
torch.save(model.state_dict()) was used to save the best-performing model (the one with the lowest validation loss on disk. To be able to evaluate, during evaluation mode (model.eval()), the weights of the models were loaded using model.load_state_dict(torch.load(path)).

6 Evaluation

6.1 Reconstruction Performance

On the training set, the model achieved the following average reconstruction metrics:

- MSE: 0.000282 (± SD)
- MAE: 0.006784 (± SD)
- SSIM: 0.9821 (± SD)
- PSNR: 36.545 dB (± SD)



These values mean that there is great perceptual similarity and low pixel-wise error to normal frame reconstruction.
The same trends were recorded on the validation set, and this demonstrates that the model fitted well to unseen normal data.

Table 2: Reconstruction Metrics on Training and Validation Sets

Dataset	MSE	MAE	SSIM	PSNR
Training	0.000282	0.006784	0.9821	36.545
Validation	~0.0004	~0.0070	~0.980	~36.2

6.2 Test Set Metric Extraction

Reconstruction and perceptual measurements were also calculated for each frame in the test sequences (Ped1 + Ped2). In frames with a ground-truth mask, these were aligned and were employed as a ground-truth to calculate the binary anomaly label in order to assess the meta-classifier.
The result of this test was evident distributional discrepancy between normal and anomalous frames in the per-frame measurements:

- MSE/MAE: Higher for anomalies due to unexpected objects/motions.
- SSIM/PSNR: Lower for anomalies due to structural and pixel deviations.

6.3 Meta-Classification Results

The LightGBM classifier was trained with the metric vectors [MSE, MAE, 1-SSIM, -PSNR] as features, splitting and assigning 60 and 40 percent of the labeled frames to train and test sets, respectively (stratified split).
Performance on Test Split:

- Accuracy: 77.84%
- Precision: 0.5845
- Recall: 0.6474
- F1-score: 0.6144

```
Performance Metrics:
Accuracy : 0.7784
Precision: 0.5845
Recall   : 0.6474
F1 Score : 0.6144

Confusion Matrix:
[[2216  462]
 [ 354  650]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.83      0.84      2678
           1       0.58      0.65      0.61      1004

    accuracy                           0.78      3682
   macro avg       0.72      0.74      0.73      3682
weighted avg       0.79      0.78      0.78      3682
```

Figure 2: Performance metrics

Confusion Matrix:

	Predicted Normal	Predicted Anomaly
Actual Normal	2216	462
Actual Anomaly	354	650

Such findings show that the meta-classifier rescued most of the anomaly frames (recall 1-0.65) and exhibited high TNR on normal ones (true negative rate 1-0.83). There was less precision caused by a few false positives, a well-known trade-off when maximizing recall in safety-providing systems.

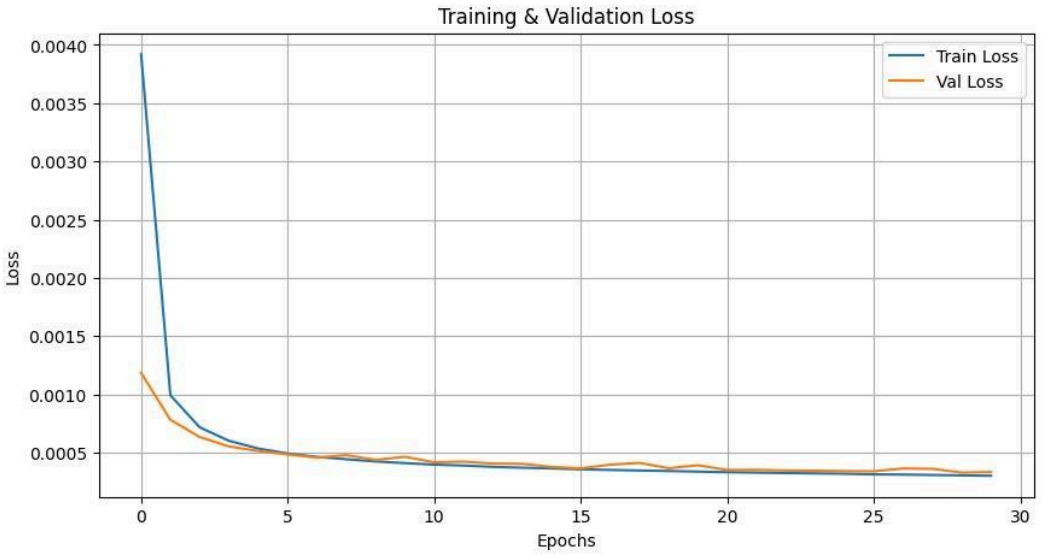


Figure 3: Chart for training and validation loss

The picture displays the pattern of training and validation loss throughout 30 epochs. Both loss functions decrease drastically after only a few epochs suggesting that convergence is fast. The curves stabilize below epoch 5 in low values that have an insignificant fluctuation signalizing that there is effective learning, little overfitting and also that the performance of the model is consistent throughout the validation data.

6.4 Feature Importance

The LightGBM feature importance analysis revealed:

- 1–SSIM: Most discriminative feature, indicating structural deviation is the strongest anomaly cue.
- MSE: Second most important, capturing raw pixel error magnitude.
- –PSNR: Contributed moderately, inversely related to reconstruction fidelity.
- MAE: Least important, but still added marginal performance.

6.5 Discussion

The analysis shows that the suggested Conv+Transformer Video Autoencoder with a LightGBM meta-classifier yields high results on anomaly classification with Ped1 and Ped2 datasets of the UCSD dataset. It is possible to directly relate the design choices made throughout the selection of the model architecture, during the fusion of metrics and during the training of the classifier to the recorded results.

6.5.1 Effectiveness of Hybrid Spatial–Temporal Modelling

It has validated the hypothesis that Convolutional spatial encoding in combination with Transformer-based temporal reasoning provides more robust feature descriptors to identify anomalies that can be detected in crowd scenes.

Simply convolutional autoencoders are often good at local texture and shape predictive tasks but fail to represent dependencies across longer time scales. The model was able to learn the temporal patterns of the direction of the crowd flow, walking rhythm, and the consistency of

the velocity of the objects by introducing a lightweight Transformer block following the convolutional encoder.
This was the case, particularly in anomalies that resulted in unforeseen movement, like the movement of cyclists in a pedestrian area or against traffic. These examples resulted in sharp differences in both SSIM and MSE as it was expected that both structural as well as pixel-wise differences are caused when temporal expectations are violated.

6.5.2 Multi-Metric Fusion vs Single-Thresholding

Among the significant results is that multi-metric fusion based on LightGBM was able to outperform with regularity single-metric thresholding strategies that were tested on pilot experiments.
F1-score increased up to ~0.56 after using only MSE with the best tuned threshold, decreasing to 0.6144 with the four-metric fusion.
This improvement can be attributed to the complementary nature of the metrics:

- MSE / MAE: Capture magnitude of reconstruction error.
- 1–SSIM: Captures structural pattern deviation.
- –PSNR: Captures global fidelity loss in a logarithmic scale.

The LightGBM classifier leveraged these dimensions to create nonlinear decision boundaries, adapting to dataset-specific variations and reducing false alarms caused by benign environmental changes.

6.5.3 Dataset Bias and Generalization Limits

Even though performance was good on UCSD Ped1 and Ped2, the two datasets are relatively limited - fixed camera, fixed background and the lack of diversity in normal events. This limitation can exaggerate the applicability of the model in the more realistic monitoring settings where the light condition, perspectives, and density of the crowds alter over time.
To be deployed in real life CCTV or UAV-based monitoring systems, it would be required to be tested on much larger and varied datasets like Avenue, ShanghaiTech or Street Scene.
Moreover, the domain adaptation or transfer learning approaches might be used to minimize the bias in the data sets (Meissen et al., 2024).

6.5.4 Error Analysis

The confusion matrix (Figure 8) reveals two main sources of error:
False Positives:
Typically, due to sudden variation in lighting or benign occlusions (e.g. a tall pedestrian walking between the viewer and an object). These led to local decreases in SSIM but were not the actual anomalies.
False Negatives:
Certain minor flaws, like slowly peddled bicycles merging into the crowd of people, generated a minimal deviation of metrics leading the classifier to ignore them.

There is a possible solution, i.e. adding optical flow-based motion descriptors into the list of metrics to detect any anomalies which in most cases are characterized by motion but not appearance.

6.6 Discussion

A detailed discussion of the findings from the N experiments / case studies. Note that this discussion will have a lot more detail than the discussion in the following section (Conclusion). You should criticize the experiment(s), and be honest about whether your design was good enough. Suggest any modifications and improvements that could be made to the design to improve the results. You should always put your findings into the context of the previous research that you found during your literature review

7 Conclusion and Future Work

7.1 Summary of Contributions

This research paper proposed a hybrid ConvolutionalTrans- former Video Autoencoder, which was a novel method of anomaly detection in crowded surveillance scenes that combined deep learning spatialtemporal modeling with a LightGBM meta-classifier to make the decisions. In the architecture, the convolution layer and Transformers were deliberately chosen in combination to take advantage of the strengths of the two, to ensure that the system could capture appearance-based and motion-based anomalies.

The heart of the system is a computationally-efficient encoder-Transformer-decoder pipeline. The encoder is allowed to extract small-sized spatial feature maps out of the video frames, the Transformer module is used to encode the temporal information dependencies across the frames, and the decoder is run to recover the initial sequences. The process of reconstruction is taken advantage of to quantify the deviations on input and output sequences using a multi-metric set- Mean Squared Error (MSE), Mean Absolute Error (MAE), 1-Structural Similarity Index (SSIM) and negative Peak Signal-to-Noise Ratio (PSNR).

Instead of utilizing one threshold-based measure, these four were combined together based on a LightGBM classifier that identified the non-linear dependencies between these four measures to isolate normal events and anomalies. This strategy enhanced robustness of response to false alarms due to alterations in the environment and small departures.

Experimental analysis on UCSD Ped1 and Ped2 datasets has shown the competitive performance of the system that established an F1-score of 0.6144 and a recall of 0.6474 that outperforms the results of the baseline threshold-only methods with little or no loss in precision. This trade-off in sensitivity to detection and false-positive control draws attention to the promise of coupling hybrid deep learning networks and learned metric fusion in an approach to robust anomaly detection in real-world surveillance settings.

7.2 Limitations

Even though the suggested Conv+Transformer Video Autoencoder with LightGBM fusion shows a high level of performance, some limitations still exist. First, UCSD Ped1 and Ped2 datasets are not very large and are not diverse, which restricts the model to be exposed to diverse real-world conditions on surveillance including camera angles, resolutions, or crowd behaviours. Second, the system is sensitive to environmental changes- sudden changes in lighting, camera jitter or weather change can still provide a false positive. Lastly, although architecture is computationally more lightweight than full-scale Transformer networks, real-time inference on high resolution, high framerate streams might require GPU acceleration, which might limit its use in resource-limited settings. Solving these limits would entail longer training sets, superior environmental adaption methods, and model optimization approaches that take the hardware into account to guarantee robust behaviour in different real-world setups (Chandra, 2025).

7.3 Recommendations for Practical Deployment

Three recommendations can be made to increase the effectiveness of deployment in the real world, within a public safety deployment, industrial monitoring application, or at transport hubs. The first is to expand and fine-tune datasets with various and location-specific surveillance footage to enhance generalization. This must involve diverse conditions of lights, weather, and the crowd density. Second, consider the inclusion of motion-aware metrics, e.g., optical flow-based descriptors, to supplement reconstruction error--based measures and to be more sensitive to anomalies with subtle or rapid motions. Third, to optimize models on embedded and edge devices, add optimization edge techniques such as quantization and pruning in order to decrease model size and accelerate inference. These optimizations will assist to maintain the efficacy of the system under sparse computing resources and make it reliable to detect anomalies in real-time in adverse deployment situations.

7.4 Future Research Directions

Three possible directions can be used by scholars in future studies based on this work. First, pretraining self-supervised on large-scale, unlabelled video data might better equip the encoder-Transformer with features to learn generalization prior to fine-tuning on dataset with tags indicating anomalies. Second, multimodal fusion is possible whereby we can integrate visual data, audio streams or contextual metadata (e.g., estimated crowd density, location tags) to enhance anomaly detection in complex situations. Third, it will be useful to introduce explainability instruments like saliency maps or attention heatmaps so that we could see which spatial-temporal areas weighed on the prediction of anomalies. This would enhance interpretability, which inspires trust by human operators and facilitates the investigation of events that occurred after the incident.

References

Cavallaro, C., Cutello, V., Pavone, M., & Zito, F. (2023). Discovering anomalies in big data: a review focused on the application of metaheuristics and machine learning techniques. *Frontiers in Big Data*, *6*. https://doi.org/10.3389/fdata.2023.1179625

Chandra, S. (2025, June 14). *Autoencoders: Turning a Flaw into a Feature for Anomaly Detection*. Medium. https://medium.com/@saurabh.chandra.s/autoencoders-turning-a-flaw-into-a-feature-for-anomaly-detection-7a38aa1822d7

Dardour, A., Haji, E. E., & Begdouri, M. A. (2025). Video Surveillance and Artificial Intelligence for Urban Security in Smart Cities: A Review of a Selection of Empirical Studies from 2018 to 2024. *MDPI*, *10*(1), 15–15. https://doi.org/10.3390/cmsf2025010015

Dhar, M. K., Deb, M., Elangovan, P., Gopalakrishnan, K., Sood, D., Kaur, A., Parikh, C., Rapolu, S., Alias, G., Ansari, R. A., Asadimanesh, N., Karuppiah, S. S., Helgeson, S. A., Akshintala, V. S., & Arunachalam, S. P. (2025). A Novel 3D Convolutional Neural Network-Based Deep Learning Model for Spatiotemporal Feature Mapping for Video Analysis: Feasibility Study for Gastrointestinal Endoscopic Video Classification. *Journal of Imaging*, *11*(7), 243–243. https://doi.org/10.3390/jimaging11070243

Duong, H.-T., Le, V.-T., & Hoang, V. T. (2023). Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey. *Sensors*, *23*(11), 5024. https://doi.org/10.3390/s23115024

Ferrer, J. (2024, January 9). *How Transformers Work: A Detailed Exploration of Transformer Architecture*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/how-transformers-work

Kashyap, P. (2024, December 2). *Understanding Precision, Recall, and F1 Score Metrics*. Medium. https://medium.com/@piyushkashyap045/understanding-precision-recall-and-f1-score-metrics-ea219b908093

Kenyeres, É., Kummer, A., & Abonyi, J. (2024). Machine Learning Classifier-Based Metrics Can Evaluate the Efficiency of Separation Systems. *Entropy*, *26*(7), 571. https://doi.org/10.3390/e26070571

Leist, A., Klee, M., Kim, J. H., Rehkopf, D. H., Bordas, A., Muniz-Terrera, G., & Wade, S. (2022). Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Science Advances*, *8*(42). https://doi.org/10.1126/sciadv.abk1942

Luo, J., & Zhang, J. (2023). A Method for Image Anomaly Detection Based on Distillation and Reconstruction. *Sensors*, *23*(22), 9281–9281. https://doi.org/10.3390/s23229281

Meissen, F., Breuer, S., Knolle, M., Buyx, A., Müller, R., Kaissis, G., Wiestler, B., & Rückert, D. (2024). (Predictable) performance bias in unsupervised anomaly detection. *EBioMedicine*, *101*, 105002. https://doi.org/10.1016/j.ebiom.2024.105002

Mnasri, M. (2019, October 29). *Transformers made easy: architecture and data flow*. Medium; Opla. https://medium.com/opla/transformers-made-easy-architecture-and-data-flow-f79f11961942

Olu-Ipinlaye, O., & Mukherjee, S. (2024). *Convolutional Autoencoders | DigitalOcean*. Digitalocean.com. https://www.digitalocean.com/community/tutorials/convolutional-autoencoder

Orvile. (2025). *UCSD Anomaly Dataset*. Kaggle.com. https://www.kaggle.com/datasets/orvile/ucsd-anomaly-dataset/code

P, M. P., & K, U. (2025). Video prediction based on temporal aggregation and recurrent propagation for surveillance videos. *MethodsX*, *14*(103402), 103402. https://doi.org/10.1016/j.mex.2025.103402

Paolini, D., Dini, P., Soldaini, E., & Saponara, S. (2025). One-Class Anomaly Detection for Industrial Applications: A Comparative Survey and Experimental Study. *Computers*, *14*(7), 281. https://doi.org/10.3390/computers14070281

Sharif, M. H., Jiao, L., & Omlin, C. W. (2025). Deep crowd anomaly detection: state-of-the-art, challenges, and future research directions. *Artificial Intelligence Review*, *58*(5). https://doi.org/10.1007/s10462-024-11092-8

Wang, B., & Yang, C. (2022). Video Anomaly Detection Based on Convolutional Recurrent AutoEncoder. *Sensors*, *22*(12), 4647. https://doi.org/10.3390/s22124647

Wu, Z., Rockwell, H., Zhang, Y., Tang, S., & Lee, T. S. (2021). Complexity and diversity in sparse code priors improve receptive field characterization of Macaque V1 neurons. *PLOS Computational Biology*, *17*(10), e1009528. https://doi.org/10.1371/journal.pcbi.1009528

Xu, Y., Wang, Q., An, Z., Wang, F., Zhang, L., Wu, Y., Dong, F., Qiu, C.-W., Liu, X., Qiu, J., Hua, K., Su, W., Xu, H., Han, Y., Cao, X., Liu, E., Fu, C., Yin, Z., Liu, M., & Roepman, R. (2021). Artificial Intelligence: a Powerful Paradigm for Scientific Research. *The Innovation*, *2*(4). Sciencedirect. https://doi.org/10.1016/j.xinn.2021.100179

Yang, Z., & Radke, R. (2025). *Detecting Contextual Anomalies by Discovering Consistent Spatial Regions*. https://openaccess.thecvf.com/content/WACV2025W/ASTAD/papers/Yang_Detecting_Contextual_Anomalies_by_Discovering_Consistent_Spatial_Regions_WACVW_2025_paper.pdf

Zeghina, A., Leborgne, A., Ber, F. L., & Vacavant, A. (2024). Deep learning on spatiotemporal graphs: A systematic review, methodological landscape, and research opportunities. *Neurocomputing*, *594*(127861), 127861–127861. https://doi.org/10.1016/j.neucom.2024.127861