





Review

The Role of Deep Learning Models in the Detection of Anti-Social Behaviours towards Women in Public Transport from Surveillance Videos: A Scoping Review

Marcella Papini ^{1,2,*} , Umair Iqbal ¹ , Johan Barthelemy ³  and Christian Ritz ¹ 

¹ SMART Infrastructure Facility, University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia; umair@uow.edu.au (U.I.); critz@uow.edu.au (C.R.)

² School of Information and Physical Sciences, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia

³ NVIDIA, 2788 San Tomas Expressway, Santa Clara, CA 95051, USA; jbarthelemy@nvidia.com

* Correspondence: marcella.papini@newcastle.edu.au

Abstract: Increasing women's active participation in economic, educational, and social spheres requires ensuring safe public transport environments. This study investigates the potential of machine learning-based models in addressing behaviours impacting the safety perception of women commuters. Specifically, we conduct a comprehensive review of the existing literature concerning the utilisation of deep learning models for identifying anti-social behaviours in public spaces. Employing a scoping review methodology, our study synthesises the current landscape, highlighting both the advantages and challenges associated with the automated detection of such behaviours. Additionally, we assess available video and audio datasets suitable for training detection algorithms in this context. The findings not only shed light on the feasibility of leveraging deep learning for recognising anti-social behaviours but also provide critical insights for researchers, developers, and transport operators. Our work aims to facilitate future studies focused on the development and implementation of deep learning models, enhancing safety for all passengers in public transportation systems.

Keywords: deep learning models; anti-social behaviour detection; women's safety; public transport; safe transportation



Citation: Papini, M.; Iqbal, U.; Barthelemy, J.; Ritz, C. The Role of Deep Learning Models in the Detection of Anti-Social Behaviours towards Women in Public Transport from Surveillance Videos: A Scoping Review. *Safety* **2023**, *9*, 91. <https://doi.org/10.3390/safety9040091>

Academic Editor: Raphael Grzebieta

Received: 25 October 2023

Revised: 1 December 2023

Accepted: 7 December 2023

Published: 13 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Public transport services are enablers of economic development. They provide individuals with access to medical facilities, workplaces, schools, and shops. However, public transport is experienced in a gendered way; mobilities shape how violence against women is perpetrated and how women experience it in public transport [1,2]. Global-, regional- and local-level studies have shown that women experience various forms of violence in public transport [3–5]. The types of violence experienced varied from demeaning remarks and threats or intimidation to robbery. The fear of violence in public spaces has the potential to affect the behaviour of passengers to the point of avoiding travelling [6].

Violent or anti-social behaviour can take many forms; there are many different definitions for anti-social behaviour, and definitions vary across contexts, social norms, and values [7]. In [8], for example, anti-social behaviour is defined as a psychological term for actions that fall outside of the realm of what is considered normative in a particular society or culture. In the present study, we use the terms anti-social and abnormal interchangeably and follow the definition of anti-social behaviour adopted by the Western Australia Police, which defines anti-social behaviour as any ‘behaviour that disturbs, annoys or interferes with a person’s ability to go about their lawful business’ [9].

Anti-social behaviour towards women in public transport is relevant to transport operators since they are concerned about improving the travel experience and ensuring customers' safety. This matter is also pertinent to governments and businesses, as closing the gender gap in economic participation could lead to organisations yielding better economic performance and outcomes [10–12].

Although video and audio surveillance technologies have the potential to address this issue, these technologies are inefficient for numerous reasons, such as the need to check the video frame-by-frame [13]. An intelligent detection system may help reduce operational issues, decreasing the transport operators' efforts and costs. Previous studies have shown that technologies based on machine learning models surpass traditional technologies in anti-social behaviour detection areas [14–18]. However, could machine learning-based detection systems be a useful and efficient technology to deal with behaviours that may influence women's perceptions of safety in public transport? To answer this question, this paper proposes a systematic scoping review that aims to understand the theoretical and practical benefits and limits to using deep learning-based technologies to detect from surveillance data anti-social behaviours associated with a threat to female commuters. Our goal is to support not only researchers and developers but also transport operators in the development and implementation of deep learning models for the detection of anti-social behaviours towards women in public transport. The scoping review will provide a timely summary of the literature describing the benefits of deep learning-based detection technologies, challenges in the automated detection of anti-social behaviours, and considerations for existing video and audio datasets that can be used to train such algorithms.

The rest of the paper is organised as follows. Section 2 presents the research methodology based on the scoping review approach adopted in this study. Section 3 describes the results of the scoping review, whereas Section 4 discusses the scoping review results obtained. Finally, Section 5 concludes the work.

2. Materials and Methods

The goal of the present study is to explore the benefits and limitations of using deep learning-based technologies to recognise anti-social behaviours threatening female commuters from audio and/or video surveillance data. For that, we adopted a scoping review methodology and followed the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) [19]. Scoping reviews are used to introduce a wide overview of the evidence relevant to a topic and are valuable when examining areas that are emerging, clarifying key ideas, and identifying gaps [20]. To undertake a scoping review, we used the methodological framework proposed in [21] and refined in [22–24], which includes the following steps: (a) *identification of the research questions*, (b) *identification of studies relevant to the research questions*, (c) *selection of studies for inclusion*, (d) *charting of information and data within the included studies*, and (e) *summarising and reporting of the review results*. These steps as well as their sub-steps and outcomes are illustrated in Figure 1 and described in the next sections.

2.1. Identification of the Research Questions

The first step of the scoping review methodology consisted of identifying the research questions, which help shape and carry out the scoping review. The research questions that the present scoping review aims to answer are as follows.

- What are the benefits of deep learning models in the detection of anti-social behaviours in public spaces?
- What deep learning models have been used to detect anti-social behaviours in public spaces?
- What audio and/or video datasets are relevant to the detection of anti-social behaviours towards women commuters?
- What are the challenges in the use of deep learning models for the detection of anti-social behaviour(s) in public transport?

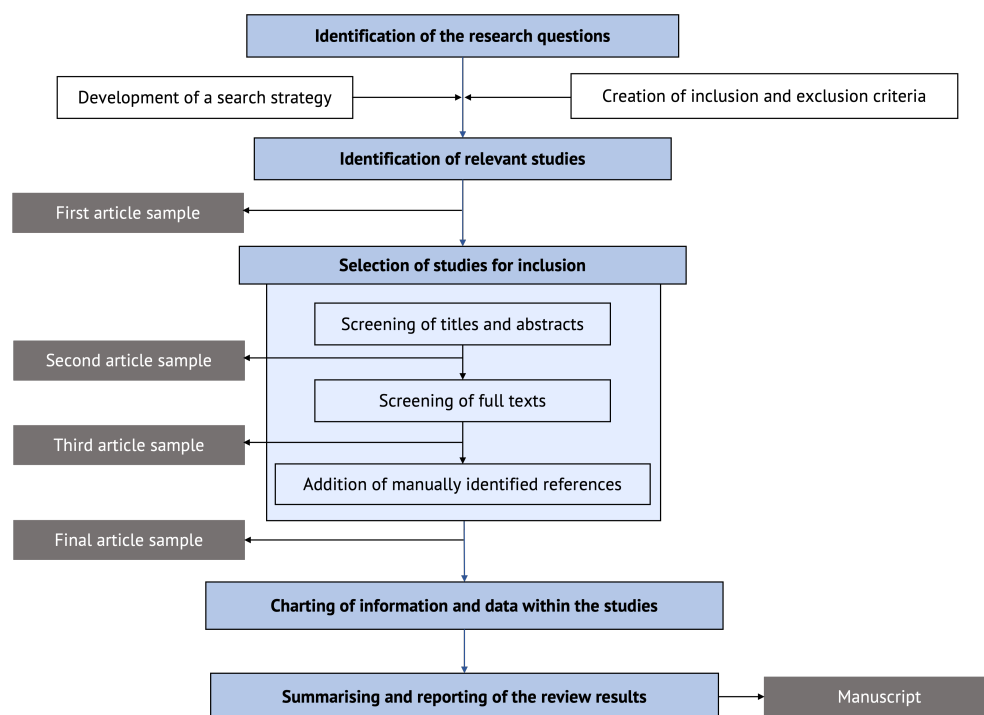


Figure 1. Study’s methodological framework.

2.2. Identification of Studies Relevant to the Research Questions

The second step of the scoping review methodology is the identification of studies that are pertinent to the identified research questions. The goal of this step is to identify relevant studies in the field of anomaly detection by conducting a comprehensive literature search. For that, this step includes two activities (see Figure 1): (a) development of a search strategy and (b) creation of inclusion and exclusion criteria. These activities are described as follows.

2.3. Development of a Search Strategy

The search strategy specifies the academic reference and citation databases and the search terms to be utilised for finding pertinent articles.

On the one hand, we chose reputable databases relevant to engineering and computer science disciplines, namely, <https://ieeexplore.ieee.org> (accessed on 1 November 2022), <https://clarivate.com> (accessed on 1 November 2022), <https://www.scopus.com> (accessed on 1 November 2022), <https://link.springer.com/> (accessed on 1 November 2022), and <https://dl.acm.org> (accessed on 1 November 2022).

On the other hand, we adopted a broad definition of keywords for search terms, as recommended in [21], to obtain a broad coverage of the available literature. The search terms used were: “anomaly”, “unusual”, “abnormal”, “anomalous”, “violent”, “harass”, “harassment”, “sexual”, “unsafe”, “panic”, “anti-social”, “threatening”, “violation” OR “violations” AND “behaviour”, “behavior”, “motion” OR “activity” AND “detection” OR “recognition” AND “audio”, “video”, “surveillance”, “sound” OR “speech” AND “public transport”, “public transportation”, “transport”, “transportation” OR “public spaces” AND “deep learning”, “machine learning” OR “neural network”. It is important to note that the words “woman” and “women” were not included as search terms. This is for the purpose of looking for examples of research focusing on similar safety issues. By broadening the focus at this search stage, it was possible to identify, as the discussion of the results shows (see Section 4), automated detection technologies that have not been used to detect anti-social behaviours perceived by women, but that are potentially suitable candidates.

2.4. Creation of Inclusion and Exclusion Criteria

After the development of a search strategy, we created inclusion and exclusion criteria, which are displayed in Table 1, to ensure the relevance of the selected papers. It was decided that all publications in English should be included, and no restrictions should be placed on the publication dates and study design. As the focus of this scoping review is on technologies for the detection of anti-social behaviours of perpetrators, only publications that fall within the anomaly detection research area should be included. Other anomaly detection research areas, such as traffic anomalies and environmental anomalies, are concerned with anti-social behaviours that either are not relevant, e.g., facial sleepiness expressions [25], or do not influence women's perception of safety in public transport, e.g., an individual abandoning an object [26]. Finally, only publications that considered anti-social behaviours that affect women's perception of safety were included in the sample.

Some of the sampled publications mentioned what anti-social behaviour could be identified by the proposed algorithm; others did not. However, based on the audio or video dataset adopted for training and/or assessing the performance of the proposed algorithm, we inferred what anti-social behaviour(s) could be detected by the introduced technology.

To decide what detected behaviour was of relevance to women commuters' safety perspective, we followed the definition of anti-social behaviour from the Western Australia Police, which was presented in Section 1.

Table 1. Inclusion and exclusion criteria.

Criteria	Inclusion Criteria	Exclusion Criteria	Justification
Language	English	Non-English language	English is the official language of the study's team
Time	All studies prior to 2022	N/A	Any deep-learning detection technology, regardless of when it was first proposed, is relevant
Study design	All study designs were included in the review	N/A	This review is concerned with the breadth of existing knowledge
Study topic	Publication falls within the scope of anomaly detection	Publication does not fall within the pertinent scope	This review is only interested in the detection of anti-social behaviours from perpetrators
Anti-social behaviour detected	Publication addresses an anti-social behaviour that could affect women's safety perception	Publication does not make clear what anti-social behaviour is studied or anti-social behaviour considered is not relevant	Only deep learning models that can detect relevant anti-social behaviours are of interest

2.5. Selection of Studies for Inclusion

The third step of the scoping review methodological framework is the selection of relevant articles for inclusion in the article sample. A PRISMA flow diagram illustrating the procedures in this step is presented in Figure 2. These procedures ensure that the review is comprehensive, relevant, and unbiased.

The studies identified through the search strategy (see Section 2.3) formed the first sample of articles (see Figure 1), which contained 92 papers. Once the first sample was obtained, we proceeded with its refinement. Accordingly, two screening processes were undertaken in this step, and the inclusion and exclusion criteria guided these processes.

To manage and organise them, the publications in the first sample were imported into Covidence, a web-based systematic review software that enables various reviewers to work more efficiently through the steps of a systematic review [27]. When the first sample was imported into the software, 16 duplicated publications were identified and then excluded, resulting in the second sample ($n = 76$). The first screening process consisted of two authors independently reviewing each title and abstract to exclude publications that did not adhere to the inclusion criteria. Papers that passed through the first screening process formed the second article sample ($n = 28$) and were then read in full by two authors independently

to assess based on the inclusion criteria whether the paper should be inserted in the third article sample. The third sample included 10 manuscripts.

Conflicts at each stage of the screening processes were resolved through researcher discussion until agreement was reached. Data extraction was completed in Covidence using a data extraction template. The information extracted is presented in Table 2.

After we acquired the third sample of articles, we manually identified 72 additional potentially relevant references from the third sample. These potentially pertinent papers also went through the first and second screening processes, and those that moved through the two processes ($n = 23$) were added to the third sample, resulting in the final sample of articles, which contained 33 manuscripts.

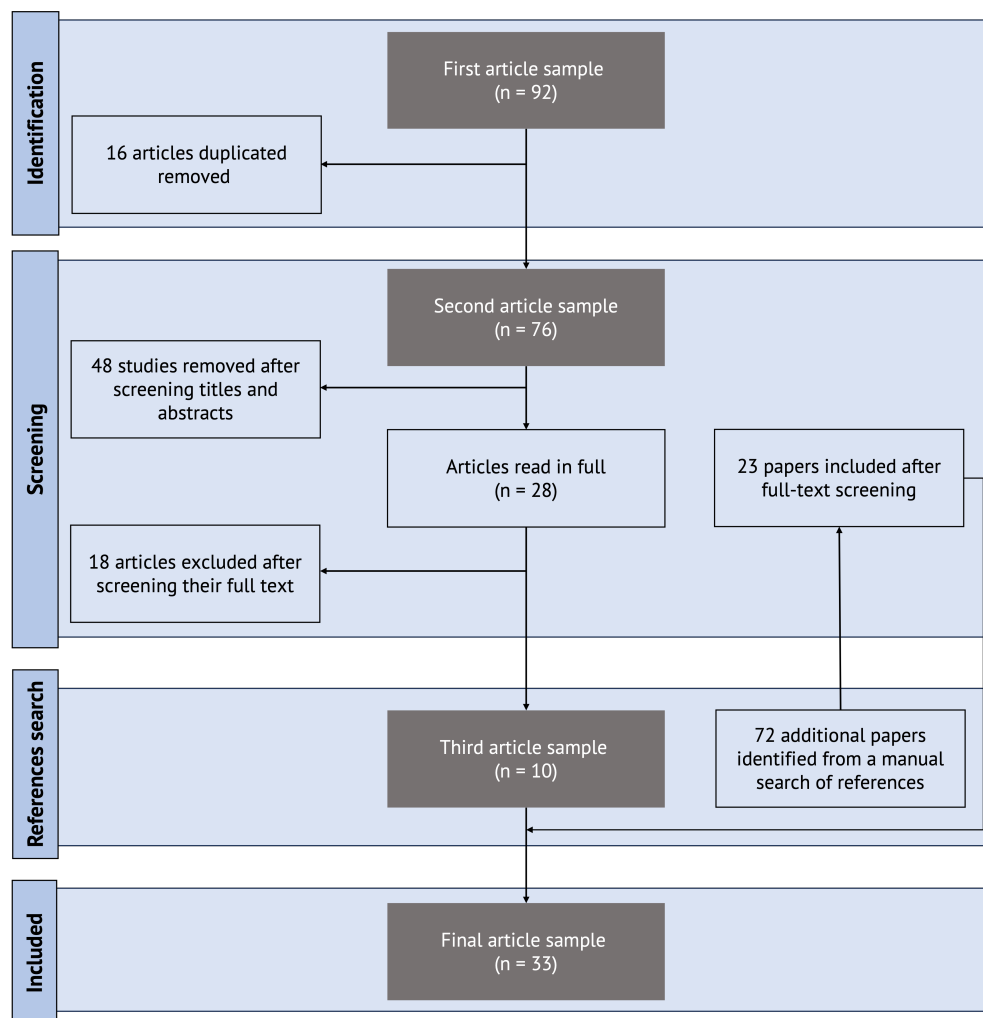


Figure 2. PRISMA flow diagram for paper selection.

Table 2. Data extraction template.

Characteristics	Information to Be Extracted
Authors, year, and aim(s)	Who were the authors of the study? When was it conducted? What was the aim of the study?
Research method	What research method did the publication follow?
Detected anti-social behaviour	What type of anti-social behaviour(s) does this publication consider? Is it an anti-social behaviour that would influence women's perceived safety in public transport settings?

Table 2. *Cont.*

Characteristics	Information to Be Extracted
Type of technology	What type of deep learning detection model was adopted/introduced in the publication? What is the technical infrastructure necessary?
Limitations and opportunities	What are the constraints related to deep learning detection technologies? What are the benefits of these technologies? What and how do environmental and operational factors limit/improve the efficacy of these technologies?
Training and validation dataset	Does the publication use an audio/video dataset to train the algorithm or to assess the performance of the technology? If so, what dataset? And is it available?

2.6. Charting of Information within the Included Studies

The last stage of the scoping review methodological framework is the charting of the information within the articles included in the final sample by summarising each paper, including the details extracted using the data extraction template (see Table 2). A summary of the sampled articles is available in Appendix A.

Among the studies in the sample, 10 are surveys or review manuscripts, while the remaining 24 are applied research where a new or improved automated detection technology are presented. Of the applied research publications, 4 studied acoustic surveillance and 20 video surveillance. All sample articles fall within the research area of anomaly detection in public spaces. From the article sample, one can presume that audio/video anomaly detection in public settings has been researched since 2007, with an increased interest in the topic from 2018 and the largest number of articles (10) recorded in 2019. However, despite the increased gender-based vulnerabilities experienced by women in public places [6], in accordance with our article sample, it seems that women's perception of safety has not been given consideration in the research area of anomaly detection. Although every applied research publication in the sample introduced an algorithm to identify one or more behaviours that could impact women's safety perspective in public transport, none of the technologies was proposed with the aim of detecting perpetrators' behaviours towards women commuters. With respect to the training datasets, some of the applied research studies adopted their own datasets, which were designed based on real data, and others utilised well-known datasets. As highlighted in Section 2.4, we used the dataset adopted in an applied research manuscript to infer what anti-social behaviours could be detected by the proposed technology when that was not specified in the study.

2.7. Summarising and Reporting the Review Results

The last step of the scoping review is summarising and reporting the review results. The results of the present scoping review are presented and discussed in the next section.

3. Synthesis of the Literature

Following the questions in the data extraction template (see Table 1), we undertook a synthesis of the literature on the automated detection of anti-social behaviours in public spaces based on the identified research questions (see Section 2.1). These questions are addressed in detail as follows.

3.1. What Are the Benefits of Deep Learning Models in the Detection of Anti-Social Behaviours in Public Spaces?

The major potential benefit of deep learning-based detection technologies presented in the literature is to improve the level of safety of passengers [7,8,13,14,28]. The primary aim of audio and/or video surveillance systems in public transport is to assist in preventing crime and terrorist activities. Detecting anti-social activities at the early stage and even

predicting these activities before they happen is of great value to prevent serious incidents from occurring [28].

The traditional procedure for the task of capturing the images and audio of possible transgressions from surveillance systems is to appoint human controllers (e.g., security guards) whose job is to analyse large volumes of repetitive and monotonous images from multiple cameras [16,29]. These exhausting human efforts make it very difficult for the controller to always remain vigilant, which might lead to abnormal events going unnoticed [15–18,29]. Moreover, if the station workers are off at night when an incident occurs, a belated response may have dangerous results. If automated detection technologies are utilised, authorities can respond to abnormal and violent behaviours sooner, maximising convenience and the feeling and the actual level of security [16,30]. Nonetheless, none of the sampled publications has validated that passengers felt safer because an automated detection technology was in place.

Another potential benefit of deep learning-based detection technologies is that they might be used for monitoring people who fall on train tracks and/or when platforms/carriages are too crowded, etc. Automated detection technologies can also be cost-effective. These technologies can reduce the costs associated with the manual monitoring incurred by transport operators [15,18,31–33]. Additionally, deep learning detection models usually only require infrastructure that is already available in public spaces, that is, a set of closed-circuit television (CCTV) cameras [34].

3.2. What Deep Learning Models Have Been Used to Detect Anti-Social Behaviours in Public Transport?

The deep learning model proposed in most of the 24 applied research articles in the sample was the convolutional neural networks (CNN) model, followed by the Gaussian mixture model (GMM). Nine sampled applied research manuscripts introduced a CNN-based model [15,29,31,33,35–39], whereas seven applied research articles presented a GMM-based model [40–43]. These models are briefly detailed as follows.

3.2.1. CNN-Based Deep Learning Models Sampled

Ramachandran and Palivela [36] proposed a framework aimed at detecting suspicious human behaviour in surveillance videos and distinguishing it from normal activities. The model utilised a CNN to extract features from optical flow slices and pre-trained the activities based on real-time data. These learned features were used to predict the type of activity and classified using a multi-class support vector machine (MSVM). The CNN is a type of deep learning model that is highly effective in analysing and processing structured grid-like data, such as images or sequential data. It is widely used for various tasks in computer vision, including image classification, object detection, and image segmentation. CNNs are inspired by the visual cortex of animals and exploit the spatial relationships present in the input data. They consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply filters to the input data, capturing local patterns and features. The pooling layers downsample the feature maps, reducing the spatial dimensions and extracting the most relevant information. The fully connected layers combine the high-level features and make the final predictions [44]. The system was designed for public places and security-sensitive environments. The performance of the system was evaluated using standard datasets and achieved 95% accuracy. The model involved pre-processing the input videos, extracting motion patterns, resizing and inputting them into the CNN for feature learning, and then using MSVM for classification. The proposed approach demonstrated high performance and accuracy, outperforming other classifiers like KNN and random forest models, while reducing false alarms in detecting suspicious behavior.

Proano et al. [35] addressed the challenge of detecting abnormal behaviours in surveillance videos and proposed an approach using computer vision and pattern recognition. They fully labelled a dataset of 16,853 videos, dividing them into segments and labelling

each segment as normal or abnormal. They then utilised a generic 3D convolutional neural network (C3D) to extract feature vectors from the segments and trained a multilayer perceptron (MLP) for classification. The MLP is a type of artificial neural network (ANN) that consists of multiple layers of interconnected nodes, called neurons. It is a feedforward neural network, where information flows only in one direction, from the input layer through the hidden layers to the output layer. The MLP is widely used for various tasks such as classification, regression, and pattern recognition. Each neuron in the MLP receives inputs from the previous layer, applies an activation function to the weighted sum of its inputs, and produces an output. The activation functions introduce non-linearities to the network, enabling it to learn and represent complex relationships in the data. The MLP learns by adjusting the weights and biases of its neurons through a process called back-propagation. It uses an optimisation algorithm, such as gradient descent, to iteratively minimise the difference between the predicted output and the desired output, known as the loss or error function [45]. The contribution of the paper included the labelled dataset, improved results compared to baseline research with an area under the curve (AUC) of 0.863, and the demonstration that a segment-labelled dataset enhances the classifier performance. The approach achieved promising results in abnormal behaviour detection and validated its performance through tenfold cross-validation. The proposed model consisted of segment labelling, feature extraction using the C3D network, and classification with the MLP.

Landi et al. [37] introduced a new approach to anomaly detection in surveillance videos, specifically focusing on real-world anomalies like burglaries and assaults. The proposed model considered the impact of locality by analysing spatiotemporal tubes instead of whole-frame video segments. The authors enriched existing surveillance videos with spatial and temporal annotations, creating the first dataset for anomaly detection with bounding box supervision in both the train and test sets, called UCFCrime2Local. The experimental results demonstrated that a network trained with spatiotemporal tubes outperforms a similar model trained with whole-frame videos. The study also revealed the robustness of the locality concept to different errors during the tube extraction phase at test time. Additionally, the model was capable of providing spatiotemporal proposals for unseen surveillance videos, expanding the dataset without requiring further human labelling. The contributions of the paper included the novel approach to anomaly detection using action tubes, the development of a trainable model for handling different locations within a video segment, and the creation of the UCFCrime2Local dataset with bounding box supervision. The experiments emphasised the significance of locality, the model's robustness, and its ability to provide weak annotations for new videos, highlighting the potential of considering locality in anomaly detection.

Liu and Ma [39] addressed the background-bias phenomenon in anomaly detection using deep neural networks. The paper investigated whether deep networks focus on learning the background rather than the essence of anomalies and proposed a solution to alleviate this bias. The researchers conducted experiments and verified that deep networks tend to rely on the background rather than the anomaly patterns. To tackle this problem, they introduced an end-to-end framework with a novel region loss that guided the network to focus on the anomalous region. They also incorporated meta learning to prevent overfitting and improve generalisation. The largest anomaly detection dataset was re-annotated with bounding boxes, and extensive experiments demonstrated the effectiveness of the proposed approach in reducing the influence of the background and outperforming other methods.

Majhi et al. [15] presented a two-stream CNN architecture for anomalous event detection in surveillance videos. The proposed model utilised both normal and anomalous videos to improve performance. The two-stream CNN is a deep learning architecture that combines spatial and temporal information for tasks such as action recognition in videos. It consists of two separate CNN streams: one that processes spatial information from individual frames and another that analyses temporal information from sequences of frames. The spatial stream is designed to capture static visual appearance and local spatial patterns

within individual frames. It takes an individual frame as input and applies convolutional and pooling layers to learn hierarchical representations of the visual content. The temporal stream, on the other hand, focuses on modelling motion information and capturing temporal dynamics over a sequence of frames. It takes a series of frames as input and processes them through 3D convolutional and pooling layers, which capture spatiotemporal patterns and motion information. The outputs from the spatial and temporal streams are fused or combined at a later stage, often through concatenation or element-wise operations, to obtain a unified representation that incorporates both spatial and temporal information. This combined representation is then used for making predictions or classifications. The two-stream CNN architecture enables the model to effectively leverage both static appearance and temporal dynamics, providing enhanced performance for tasks that involve analysing videos [46,47]. A database pre-processing technique was introduced to capture spatial and temporal information for each second of a video, which was then fed as input to the two-stream CNN architecture. The model was evaluated using the UCF-crime dataset and achieved a superior classification accuracy compared to other state-of-the-art techniques. The proposed method achieved 88.74% accuracy on the UCF-crime dataset with a frame processing time of 346 milliseconds. The contributions of the paper included the database pre-processing technique and the efficient two-stream CNN architecture for real-world anomalous event detection. The study suggested the potential for future enhancement of system performance using other deep learning models.

Affonso et al. [29] presented a method for detecting anomalies, specifically assaults, in the public transportation environment using CNN and CCTV cameras. The challenges in this task included equipment standardisation, low image quality, poor camera positioning, and an imbalanced dataset. The proposed method utilised CNN architectures to classify CCTV images and determine the probability of a robbery occurrence. The method achieved promising results despite the limitations of low-quality images and the absence of temporal information analysis. Future work involved exploring the impact of temporal information on the precision of the proposed method.

Ullah et al. [31] presented an efficient deep learning-based framework for intelligent anomaly detection in surveillance networks. The proposed model utilised a pre-trained CNN to extract spatiotemporal features from a sequence of frames, followed by a multi-layer bi-directional long short-term memory (BD-LSTM) model for classification of anomalous and normal events. The BD-LSTM is a variant of the LSTM recurrent neural network (RNN) architecture. It is designed to effectively capture both past and future dependencies in sequential data by processing the input sequence in both forward and backward directions simultaneously. LSTMs are a type of RNN that excel at handling sequential data due to their ability to capture long-term dependencies and mitigate the vanishing gradient problem. Bi-LSTMs extend the capabilities of traditional LSTMs by incorporating information from both the past and the future, making them suitable for tasks such as sequence labelling, sentiment analysis, and machine translation. In a BD-LSTM, the input sequence is processed in two separate LSTM layers: one in the forward direction and the other in the backward direction. Each LSTM layer consists of memory cells and gates that regulate the flow of information. By combining the outputs of the forward and backward LSTMs, the BD-LSTM captures a comprehensive representation of the input sequence, taking into account both the past and future context [48]. Extensive experiments on benchmark datasets demonstrated the effectiveness of the framework, achieving a significant increase in accuracy compared to state-of-the-art methods. The contributions of the work included the development of an adaptable framework for real-world surveillance scenarios, the utilisation of deep CNN features combined with BD-LSTM for improved accuracy, and the demonstration of real-time anomaly detection with a reduced model size and processing time.

Dileep et al. [33] introduced a real-time suspicious human activity recognition system using CNN and 2D pose estimation. The system extracted skeletal images from video frames using pose estimation to identify human poses, which were then classified using

a pre-trained CNN to determine suspicious activities like trespassing or falls. The system can generate alerts through various means to prevent unusual activities and can be applied in public places, homes, hospitals, and other surveillance areas. The proposed model combined pose estimation and CNN architecture to recognise suspicious activities accurately. The model focused on fall detection and trespassing activities and was tested on a custom dataset.

3.2.2. GMM-Based Deep Learning Models Sampled

Valenzise et al. [43] presented an audio-based video surveillance system that automatically detects anomalous audio events, such as screams or gunshots, in a public square and localises the position of the sound source to steer a video camera accordingly. The system employed two parallel GMM classifiers trained on different features to discriminate screams and gunshots from noise. The position of the sound source was estimated using time difference of arrivals (TDOA) at a microphone array and a linear-correction least square localisation algorithm. The experimental results demonstrated a high precision in event detection and accurate source localisation. The proposed approach differed from previous works by emphasising the phase of feature selection and providing insights into camera zooming based on localisation confidence. The system was planned for real-time implementation in a public square.

Clavel et al. [41] introduced a model for automatic emotion recognition in speech, focusing specifically on fear-type emotions during abnormal, life-threatening situations. The proposed model utilised the SAFE corpus, a collection of 7 h of audiovisual sequences from fiction movies containing recordings of both normal and abnormal situations. The corpus provides a wide range of emotional manifestations and addresses the lack of corpora illustrating strong emotions. The annotation strategy described the emotion and situation evolution in context. The emotion recognition system was based on dissociated acoustic models for voiced and unvoiced speech, which were merged during the classification step. The results showed promise, with an error rate of about 30%. The paper further discussed the integration of emotion recognition into automatic surveillance systems and its potential role in understanding human behaviour and diagnosing abnormal situations. The model development involved acquiring emotional data, manually annotating emotional content, describing the acoustic features, and developing machine learning algorithms. The study represented preliminary work in the emerging field of emotion recognition, particularly in the context of audio-video surveillance, with the aim of addressing the challenges posed by heterogeneous, noisy data, and specific emotional classes.

Ntalampiras et al. [40] presented an efficient methodology for acoustic surveillance of atypical situations, specifically focusing on the recognition of specific sound events such as screams, explosions, and gunshots. The proposed model utilised a probabilistic hierarchical scheme based on GMM and carefully selected sound parameters. Notably, the model included a model adaptation loop for adaptability to different sound environments. The GMM is a probabilistic model that represented a probability distribution as a combination of multiple normal distributions. It is usually used for the modelling of complex data distributions that cannot be described accurately with a single Gaussian distribution. In the GMM model, the underlying assumption is that the observed data points are generated from a mixture of K Gaussian distributions, where each Gaussian distribution in the mixture represents a distinct component or cluster in the data. The GMM assigned a probability to each data point, indicating the likelihood of it belonging to each component. This allows for the modelling of data points that may come from different sources or follow different patterns [49]. Extensive experimentation and testing, including real-world installation and operational detection rates over three days, demonstrated the effectiveness of the system in terms of recognition accuracy, miss probability, and false alarm rates. The methodology was applied to various environments, including metro stations, urban areas, and military settings. The paper also highlighted the integration of visual and infrared sensors for enhanced detection of hazardous events.

Ntalampiras et al. [42] explored the use of novelty detection in acoustic surveillance of abnormal situations. The proposed model aimed to identify unknown or novel audio data that deviated significantly from the trained data. A multi-domain feature vector was constructed using various acoustic parameters to capture diverse characteristics of audio signals. Three probabilistic novelty detection methodologies were employed and evaluated using real-world recordings from different locations with normal and abnormal sound events. The results demonstrated the effectiveness of probabilistic novelty detection in accurately identifying abnormal audio events. The paper proposed a framework that utilised multi-domain audio descriptors and achieved a high detection accuracy. Proposed future steps involved combining the acoustic component with other modalities, such as CCD and IR cameras, to improve accuracy and facilitate human behaviour detection and interpretation. The experiments aimed to evaluate the methodology further and identify any limitations.

In conclusion, various studies focusing on anomaly detection from surveillance in different contexts are presented. These studies explored the use of advanced techniques to address the challenges associated with recognising specific sound events and detecting abnormal behaviours in surveillance videos. The proposed models leverage features extracted from audio, optical flow, skeletal images, or spatiotemporal tubes, and applied classification algorithms to accurately differentiate between normal and abnormal events. The studies highlight the importance of considering spatial and temporal information, exploiting locality, addressing background bias, and incorporating emotion recognition for a comprehensive understanding of abnormal situations. These advancements contribute to improving the effectiveness of surveillance systems, enhancing public safety, and enabling early intervention in potentially hazardous situations.

3.3. What Audio and/or Video Datasets Are Relevant to the Detection of Anti-Social Behaviours towards Women Commuters?

Studies carried out in anomaly detection have proposed many training and validation datasets that are suitable for different applications. With respect to the detection of anti-social behaviours relevant to the present study, five training and validation datasets were selected from the literature. These datasets are explained in detail below and a summary of their main features is presented in Table 3.

BEHAVE

The BEHAVE video dataset, proposed in [50], includes videos with crime-oriented abnormal behaviour that were recorded in the real world using a commercial tripod-mounted camcorder. This dataset was adopted in five manuscripts in the sample. It has around 90,000 frames of humans identified with bounding boxes, providing ground-truth tracking information along with descriptions of behaviours for interacting groups. This dataset covers four video clips recorded at 25 frames per second with a resolution of 640×480 pixels. In total, there are 125 different people in this dataset, having a total of 83,545 bounding boxes for each interacting person. BEHAVE was annotated at the level of individual bounding boxes and frame-by-frame behaviour to segregate abnormal and normal events. The anti-social behaviours included in this dataset that are relevant to this study are chasing, fighting, following, and running together. No training and test partitions are provided in this dataset.

ShanghaiTech Campus

This video dataset was first proposed in [32] and was adopted in 11 manuscripts in the sample, making it the most used. This dataset was created using multiple surveillance cameras with different view angles installed at different spots to capture real events at a university campus. ShanghaiTech has challenging light conditions and is a very large dataset, containing 13 scenes, 130 abnormal events, and over 270,000 training frames. The resolution of each video frame is 480×856 pixels. For training and validation, this dataset has 330 training videos and 107 testing ones. The abnormal behaviours included in this

dataset that are relevant to this research are fighting, throwing objects, chasing, brawling, and pushing. The training set contains only normal videos, while the test set contains a large number of normal and abnormal videos, which might lead to a low performance in anomalous behaviours detection. Additionally, the anomalies in the test set are annotated at the pixel level [51].

UCF-Crime

UCF-Crime, which is a large-scale video training dataset introduced in [52], was used in seven manuscripts in the sample. The authors generated this dataset from videos on YouTube and LiveLeak using text search terms with slight variations of each anomaly term, e.g., “car crash” and “road accident”. It includes 13 anomalies with a high impact on public safety, and the anti-social behaviours pertinent to our research are abuse, arrest, assault, burglary, shooting, stealing, fighting, burglary, robbery, and vandalism. UCF-Crime consists of 1900 untrimmed surveillance videos, of which half contain anomalies, and some of the videos have multiple anomalies. The dataset is divided into two parts: the training sample, consisting of 800 normal and 810 anomalous videos, and the testing data sample, including the remaining 150 normal and 140 anomalous videos [31]. Unlike the static backgrounds in ShanghaiTech, UCF-Crime consists of complicated and diverse backgrounds, which may make it difficult to detect violent behaviours [53].

XD-Violence

XD-Violence is an audio and visual violence dataset proposed in [54]. This dataset was utilised in two manuscripts in the sample. To create this dataset, the authors collected 91 videos from both movies and YouTube by using text search queries. Violent movies were used to collect both violent and non-violent events, and non-violent movies were only used to collect non-violent events. XD-Violence is the largest-scale anomaly dataset, with a total of 217 h and 4754 untrimmed videos, consisting of 2405 violent videos and 2349 non-violent videos. This dataset includes six anomaly events and four are relevant to this research, namely, abuse, fighting, riot, and shooting. The dataset was split into two parts: the training set, containing 3954 videos, and the test set, including 800 videos, where the test set consists of 500 violent videos and 300 non-violent videos. The training set contains video-level annotations, while the test set contains frame-level annotations [51]. By containing audio-visual signals, this dataset can leverage multi-modal information and provide more confidence in the detection algorithm’s outcomes.

SAFE Corpus

The Situation Analysis in a Fictional and Emotional (SAFE) Corpus is an audio and video training dataset introduced in [41] and adopted in one manuscript in the sample. It consists of 400 audio-visual sequences from 8 s to 5 min in English generated from a collection of 30 movies from various genres: thrillers, psychological drama, horror movies, and movies that aim at reconstituting dramatic news items or historical events or natural disasters. A sequence is a movie section illustrating one type of situation, e.g., kidnapping, physical aggression, or flood. A total of 7 h and 5275 segments is obtained in which speech represents 76% of the data, and of which 71% depicts abnormal situations with fear-type emotional manifestations, among other emotions. The anomalies illustrated in the dataset that are of relevance to this study are physical/psychological threats and aggression against human beings. The speech data include about 400 different speakers. The distribution of speech duration according to gender is 47% male speakers and 31% female speakers. The remaining 20% of the spoken duration consists of overlaps between speakers, including oral manifestations of the crowd (2%). The annotation of a given segment was performed by a human labeller and influenced both by audio and video information contained in the whole sequence.

Table 3. Summary of datasets.

Dataset	Description	Relevant Anti-Social Behaviour
BEHAVE [50]	90,000 frames of humans identified by bounding boxes, 4 video clips in either WMV videos or 76,800 individual frames, recorded at 25 fps with resolution 640×480 pixels	Chase, fight, following, and running together
ShanghaiTech Campus [32]	130 abnormal events, 13 scenes integrating complex light conditions and camera angles, over 270,000 training frames, 330 training videos and 107 validation videos, resolution 480×856 pixels for each video frame	Fighting, throwing objects, chasing, brawling, and pushing
UCF-Crime [52]	128 h of videos, 1900 untrimmed videos of real-world surveillance footage extracted from the internet, with an average length of 4 min each, including 13 types of anomalous events. Before computing features, each video could be re-sized to 240×320 pixels and the frame rate fixed to 30 fps	Abuse, arrest, assault, burglary, shooting, stealing, fighting, burglary, robbery, and vandalism
XD-Violence [54]	4754 untrimmed videos with audio collected from both films and YouTube, split into 2405 violent videos and 2349 non-violent videos. Six common types of violence are covered, with a total of 217 h	Abuse, fighting, riot, and shooting
SAFE Corpus [41]	400 audio-visual sequences in English, including fear-type emotions, 5275 segments corresponding to a total of 6 h of speech organized in sequences from 8 s to 5 min long, containing about 400 different speakers	Physical/psychological threats and aggression against human beings

These datasets have two good traits. Firstly, they are large-scale datasets, which is beneficial for training generalisable methods for abnormal behaviour detection. Secondly, they include a diversity of scenarios. In this way, abnormal behaviour detection models can respond to complicated and diverse environments and be more robust, and their outcomes can be free of bias. However, most of these datasets were created from synthetic data, that is, data artificially manufactured. One of the limitations of synthetic data is that the images generally do not present a uniform pixel quality and having dataset images with the same consistent pixel quality improves the model's performance. Another limitation is that synthetic data may not accurately reflect real-world situations and may not be representative of the underlying data distribution. Accordingly, synthetic datasets may not capture the complexity and variability of real-world data, resulting in deep learning models that generate erroneous behaviour prediction on real-life data [55,56].

3.4. What Are the Challenges in the Use of Deep Learning Models for the Detection of Anti-Social Behaviour(s) in Public Transport?

Challenges in anomaly detection from audio and/or video surveillance data were a topic covered in most of the publications in the sample. These challenges are due to many factors, such as environmental conditions and the nature of the technology, and can affect the performance (accuracy) of a deep learning model. A deep learning-based detection model should be able to stably and continuously output the correct result, which is the fundamental requirement in abnormal behaviour detection [28]. While a false positive is when the target behaviour did not occur, but the detection model predicted it happened, a false negative is when the target behaviour occurred, but the detection model did not predict it. False positives and false negatives can both result in wrong decisions and lead to severe consequences [17,28]. To avoid incorrect alerts, the accuracy of the detection model must be maximised [28,57]. We classify the identified challenges based on their factors and describe them as follows.

Ambiguity

This challenge refers to the ambiguous nature of anti-social activities. In real-world scenarios, the boundary between abnormal and normal behaviours is unclear. Humans can easily recognise abnormal or typical events based on common sense, but deep learning models need to use features learned from the differences between data representing abnormal and normal behaviours to detect these events [16,17,31]. For example, a deep

learning-based model could classify two friends playing by chasing each other as anti-social behaviour. Additionally, some normal data samples may present an abnormal feature, which will reduce the detection accuracy of the model [34,58].

Background

This challenge relates to the environment where surveillance systems operate. Since this environment varies in illumination conditions over time, deep learning-based detection techniques might have difficulty detecting abnormal behaviours in some locations/time periods [16,17,31,34]. Another issue is when the background is dynamic and complex, for instance, when many people are walking through the cameras all the time. Space, video resolution, and target occlusion might limit the detection accuracy as well [15,17,34,40,59]. If the detection model was only trained in a specific space, the detection accuracy of the model might be lower when the abnormal behaviour happens in a space different from that in the training dataset. With respect to video resolution, the detection model is not able to work well when the resolution of the surveillance video is different from that of the training dataset. Target occlusion means that the target activity is blocked by objects or individuals. Consequently, the detection model cannot detect the target behaviour, and then the performance of the algorithm decreases. Additionally, the poor positioning of CCTV cameras, including camera angles, can substantially impact the performance of detection models, as the appearance of behaviours may change depending on the different positions and angles of the cameras [29,34,59,60]. In some cases, the same individual at different poses can be detected by humans with ease. However, it might be difficult for deep learning models to detect and track the same target with various poses [61].

Data Imbalance

This challenge indicates the imbalance between data on abnormal and normal events. Although normal events happen every day, the frequency of anti-social events is low, which leads to a data imbalance [28,34,38,61]. A data imbalance makes deep learning models difficult to train, as these models must analyse a set of data containing abnormal and normal events to learn to detect the target behaviour(s). Hence, if the training and validation datasets are imbalanced, it becomes arduous not only to train the detection model but also to evaluate its performance [28]. Moreover, the anti-social events may not be recorded, and a deep learning model may not be able to detect an anti-social behaviour from surveillance videos if the behaviour is not present in the training dataset [62].

Dependency and Diversity

This challenge concerns both the contextual dependency of the definition of anti-social behaviours and their diverse nature. Although there are numerous studies on abnormal behaviour detection, there is not a unified definition of abnormal behaviour [28,58], and some of the definitions introduced are dependent on the context, and thus, cannot be adopted in other abnormal behaviour detection tasks. Even the same abnormal events are unlikely to have the same characteristics and they may vary in different backgrounds. The contextual dependency of abnormal behaviours makes the detection models not adaptable [58]. Moreover, anti-social behaviour is diverse. Its scope is wide and not limited to a specific behaviour; it rather refers to a wide class of behaviours. Therefore, there is a challenge in pre-defining the structure or class of anti-social events [40,61].

Data Quality

Another challenge is dealing with audio or video data with varying levels of quality [34,58]. Audio or video data recorded from the same equipment can have various levels of quality due to variation or malfunction of the equipment, poor lighting conditions or other environmental characteristics affecting video recording quality, background noise affecting audio recording quality, or high data compression leading to loss of resolution or decoding artefacts.

Privacy and Availability

This challenge refers to data privacy limitations and the consequent lack of available training and testing datasets. During the construction of surveillance systems for transportation networks, there are certain rules issued by the government or relevant authorities that must be followed. They specify the needs and the justification of the CCTV system. These rules cover the legislative and privacy limitations [28,63]. One of the main ethical concerns in the domain of CCTV is privacy. Video analytics could identify an individual's interactions with other individuals and objects. The individual privacy of facial and behavioural information in surveillance might be compromised, especially if the data are open [28,58]. This leads to a lack of open-source datasets from real-life cases [17,34,58,61]. Furthermore, surveillance data are considered sensitive and relevant to confidential and legal issues, which restricts data sharing [30].

Uncertainty

Another challenge relates to the changes that can be made to video or image data before they are input to the detection model and affect the performance of the model. These changes are called adversarial attacks [15,31,36,64] and could be, for instance, rotating the image, adding white noise to the image, and/or changing the scale of the image. These perturbations are usually too small to be perceptible for deep learning models [28] and could significantly disrupt security systems such as surveillance systems. For example, adversarial attacks can be implemented such that an anti-social activity is not detected by the model, and then no action is taken to address/report the incident.

Trade-offs

A balanced trade-off between real-time processing and the desired level of accuracy is needed in the application of deep learning models for detecting anti-social behaviours [34,59]. The high accuracy of the detection and localisation of abnormal behaviours by deep learning-based methods is achieved at the cost of high computational complexity and a long processing time [34]. A few important factors directly relevant to achieving real-time performance include the model architecture, hardware accelerators, data pre-processing, and threshold setting. Deep learning models usually consist of many hidden layers with millions of trainable parameters, which requires substantial computational resources, making them very challenging to deploy in real time. Model optimisation techniques (e.g., model compression, model pruning) are one of the common approaches adopted to reduce the model complexity without compromising the model accuracy by a significant margin. Furthermore, at the hardware level, GPU-enabled chips are used as hardware accelerators to support the deep learning models' deployment in real time. In terms of input data, a pre-processed data feed rather than a raw data feed is often used to improve the real-time performance.

4. Discussion

Although computer vision has evolved in the last decade as a key technology for numerous applications replacing human supervision, such as abnormal behaviour detection, our article sample shows that, so far, no academic research on abnormal behaviour detection has proposed a deep learning-based model to detect anti-social behaviours towards women commuters. Nevertheless, all sampled articles considered one or more behaviours that are of relevance to this research, such as abuse [15,31,35,38,65] and assault [35,37,38,53,65].

Some of the sampled applied research articles did not mention what abnormal behaviour is detected by the proposed deep learning model, but they described what dataset was used to train/validate the model. As such, we could infer what anti-social behaviours can be detected by the introduced model. It is important to mention that the sampled articles that clarified the anti-social behaviour targeted by the introduced technology did not specifically define the behaviour. For instance, a targeted behaviour in some of the sampled articles was abuse. Abuse has a broad definition and can come in many forms, such as physical abuse, psychological abuse, and verbal abuse. Nevertheless, none of the sampled publications that considered abuse as a target behaviour introduced a definition for it. We believe that this is such because all sampled

articles are concerned with technical aspects rather than social considerations. More importantly, no sampled applied research manuscript described the patterns/features in the dataset that were looked for by the introduced deep learning model to identify the target behaviour(s), which leaves readers only guessing how to train their own algorithms for detecting the same behaviour(s) and using the same training dataset.

Despite that, this scoping review presented deep learning models that are potentially good candidates for the detection of anti-social behaviours against women while using public transport. They can therefore be used as benchmarks when developing a deep learning algorithm for that task. However, taking into consideration the ambiguity [16,17,31,34,58] and dependency and diversity challenges [28,40,58,61], before using algorithms from the literature as benchmarks, it is necessary to specify what the target anti-social behaviours are, their visual and audio features, and in what public transport setting the behaviours will be detected, e.g., on a bus or at a platform.

Other challenges in the detection of anti-social behaviours in public spaces identified from the sampled publications were the background [15–17,29,31,34,40,59–61], data imbalance [28,31,34,38,61,62], data quality [34,58], privacy and availability [17,28,30,34,58,61,63], uncertainty [15,28,31,36,64], and trade-offs [34,59]. These challenges are due to many factors, such as environmental conditions and the nature of anti-social behaviours, and they can be overcome. For instance, the imbalance between abnormal and normal events present in a dataset could be addressed by transferring data from different cities, enriching the anti-social behaviour records, and helping the deep learning algorithm discover and model the common patterns [28].

This scoping review also introduced other relevant themes from the literature on abnormal behaviour detection, namely, the benefits of automated detection technology and considerations on audio and video training datasets, and reviewed the available audio and video training datasets relevant to the detection of anti-social behaviours towards women commuters from surveillance data. There are four training and testing datasets that contain anti-social behaviours relevant to this research. The advantages of these datasets are that they are large-scale and present a range of scenarios. However, three of them consist of synthetic data. From an algorithm development perspective, the disadvantage of synthetic data is that firstly, images in the synthetic dataset usually do not have consistent pixel quality, and having dataset images from the same capturing device(s)—with the same consistent pixel quality—significantly impacts the model's performance. Secondly, real-world data define the scope of the problem very well; that is, they are considered to be more representative of the actual condition. Therefore, when implementing a deep learning model for the detection of target anti-social behaviours, it is important to develop a training and testing dataset from real data from the actual use case.

5. Conclusions

Anti-social behaviour impacts public transport satisfaction and is among the key factors why travellers choose not to use public transport. Studies have highlighted the various forms of violence experienced by women commuters. Accordingly, predicting and preventing anti-social behaviours towards women within public transport environments is imperative to enhance their safety.

This systematic scoping review aimed to comprehensively understand the practical and theoretical implications of utilising deep learning algorithms for identifying abnormal behaviours towards women commuters using surveillance data. Our study successfully addressed the identified research questions.

5.1. Existing Landscape and Gap Analysis

This study presented a limited number of proposed deep learning models specifically designed for detecting anti-social behaviours in public spaces, despite substantial advancements achieved in other domains. Notably, this review identified the absence of automated technology focusing on behaviours impacting women passengers' perceived safety. Therefore, omitting the search terms "woman" and "women" facilitated the identification of

deep learning models that have not been applied to detect behaviours relevant to women but are potentially suitable candidates.

5.2. Potential of Deep Learning Models and Identified Challenges

While our findings suggest the potential utility of deep learning models in detecting behaviours directed at women commuters, several challenges must be addressed. The discussion of the review results highlights these challenges, emphasising the need for further research. Specifically, future studies should explicitly delineate the patterns within training and validation data sets that model the target behaviours. Moreover, they should detail innovative approaches to overcoming common challenges like ambiguous data and data imbalance in this domain.

5.3. Proposals for Stakeholders

The current findings underscore the potential for transport operators to harness deep learning models in identifying undesirable behaviours from CCTV video footage. Nevertheless, the efficacy of these models depends on several critical factors, including the quality of the video footage, the accuracy of the training data, and the complexity of the behaviours targeted for detection. Privacy concerns can impede access to real-world case data, thereby potentially impacting the models' effectiveness. Consequently, researchers and developers must consider these factors when proposing novel deep learning solutions aimed at detecting anti-social behaviours affecting women commuters. Furthermore, there exists an opportunity for researchers, serving as catalysts for progress and innovation, to propose innovative solutions that overcome these challenges, thus contributing significantly to the enhancement of public safety within transport environments.

5.4. Limitations of the Present Study

Although this study benefits many stakeholders in the domain of deep learning model applications, it is important to recognise some of its limitations. Despite rigorous search strategies, certain databases, articles in non-English languages, or unpublished studies may have been excluded from the article sample, leading to potential gaps in the review. Additionally, due to its nature, the present scoping review may not delve deeply into specific research questions or provide definitive answers to focused queries.

Author Contributions: Conceptualisation, M.P.; methodology, M.P.; formal analysis, M.P.; investigation, M.P., U.I., J.B. and C.R.; writing—original draft preparation, M.P. and U.I.; writing—review and editing, M.P., U.I., J.B. and C.R.; funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from the iMove Cooperative Research Centre, with support from Transport for New South Wales and the Department of Transport and Planning Victoria (Australia).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors thank Yan Qian for the helpful discussions about this project.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Summary of literature.

Title	Year	Research Type	Surveillance Type	Technology Adopted	Relevant Behaviour Detected	Datasets Referred to/Adopted
A Review on State-of-the-Art Violence Detection Techniques [66]	2019	Review	Video	N/A	N/A	SBH Kinecr Interaction, Hockey, Movies, KARD, Media Eval, UCF 10
A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework [32]	2017	Applied Research	Video	Stacked Recurrent Neural Network	Chasing and brawling	UCF-Crime, ShanghaiTech Campus, CUHK Avenue, UCSD, Subway
A survey of video violence detection [16]	2021	Survey	Video	N/A	N/A	Hockey, Violent Flow, Movies, BEHAVE, and RWF-2000
Abnormal Behavior Detection: A Comparative Study of Machine Learning Algorithms Using Feature Extraction and a Fully Labeled dataset [35]	2019	Applied Research	Video	Generic 3D CNN, Multilayer Perceptron	Abuse, arrest, assault, burglary, fighting, robbery, stealing and vandalism	UCF-Crime
An Adaptive Framework for Acoustic Monitoring of Potential Hazards [40]	2009	Applied Research	Audio	Gaussian Mixture Model (GMM)	Screams	BBC Sound Effects Library, Sound Ideas Series 6000, Sound Ideas: the art of Foley, Best Service Studio Box Sound Effects, TIMIT, and sound effects from various internet sources

Table A1. Cont.

Title	Year	Research Type	Surveillance Type	Technology Adopted	Relevant Behaviour Detected	Datasets Referred to/Adopted
An intelligent system to detect human suspicious activity using deep neural networks [36]	2019	Applied Research	Video	CNNs, multi-class support vector machine	Fight, boxing, robbery, and pickpocket	BEHAVE, Crowd Violence, KTH, FIRE
Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey [61]	2021	Survey	Video	N/A	N/A	UCF-Crime, Avenue, PETS2009, ShanghaiTech Campus, UCSD, UMN, BEHAVE
Anomaly Locality in Video Surveillance [37]	2019	Applied Research	Video	A tube extraction module, a 3D CNN model, and a regression network	Assault, burglary, and robbery	UCF-Crime
AnoPCN: Video Anomaly Detection via Deep Predictive Coding Network [67]	2019	Applied Research	Video	Deep Predictive Coding Network, termed AnoPCN	Fighting, chasing, and pushing	ShanghaiTech Campus, CUHK Avenue, and UCSD
Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data [38]	2022	Applied Research	Video	CNN model, bi-directional long short-term memory layer	Assault and abuse	UCF-Crime and RWF-2000
BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection [18]	2020	Applied Research	Video	Bidirectional multi-scale aggregation networks	Running with panic and fighting	UCSD, UMN, CUHK Avenue, ShanghaiTech

Table A1. Cont.

Title	Year	Research Type	Surveillance Type	Technology Adopted	Relevant Behaviour Detected	Datasets Referred to/Adopted
CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks [31]	2021	Applied Research	Video	CNN and multi-layer Bi-directional Long Short-term Memory models	Fighting and abuse	UCF-Crime
Deep Anomaly Detection for In-Vehicle Monitoring-An Application-Oriented Review [17]	2022	Review	Video	N/A	N/A	UMN, UCSD, CUHK Avenue, ShanghaiTech Campus, IITBCorridor, UCF-Crime, XD-Violence, UBnormal, SVIRO-Uncertainty
Deep learning approaches for video-based anomalous activity detection [59]	2019	Review	Video	N/A	N/A	UCSD, UMN, Live Video, Avenue, Anomalous Behavior, PETS' 09, VIOLENT-FLOWS, Weizmann, ShanghaiTech Campus, CAVIAR, BEHAVE, MIT Traffic, Subway Entrance and Exitm, i-Lids bag and vehicle detection challenge
Exploring Background-bias for Anomaly Detection in Surveillance Videos [39]	2019	Applied Research	Video	3D-CNN, Meta Learning	Stealing	UCF-Crime

Table A1. Cont.

Title	Year	Research Type	Surveillance Type	Technology Adopted	Relevant Behaviour Detected	Datasets Referred to/Adopted
Fear-type emotion recognition for future audio-based surveillance systems [41]	2008	Applied Research	Audio	GMM	Kidnapping, physical aggression, fear-stress, terror, anxiety, worry, anguish, panic, distress, anger, sadness, disgust, suffering, deception, contempt, shame, despair, and cruelty	SAFE Corpus
Future Frame Prediction for Anomaly Detection - A New Baseline [68]	2018	Applied Research	Video	U-Net, FlowNet	Loitering and fighting	CUHK Avenue, USCD, and ShanghaiTech Campus
Graph Embedded Pose Clustering for Anomaly Detection [69]	2020	Applied Research	Video	Pose graphs and a Dirichlet process mixture	Fighting and throwing objects	ShanghaiTech Campus
Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection [70]	2019	Applied Research	Video	Memory-augmented autoencoder	Fighting and chasing	UCSD, CUHK Avenue, and ShanghaiTech Campus
Motion-Aware Feature for Improved Video Anomaly Detection [65]	2019	Applied Research	Video	Multiple Instance Learning (MIL)	Abuse, assault, burglary, fighting, robbery, stealing, shoplifting, and vandalism	UCF-Crime
Multi-Channel Generative Framework and Supervised Learning for Anomaly Detection in Surveillance Videos [62]	2021	Applied Research	Video	Multi-channel framework based on four Conditional GANs	Throwing objects and loitering	CUHK Avenue, USCD, and ShanghaiTech Campus

Table A1. Cont.

Title	Year	Research Type	Surveillance Type	Technology Adopted	Relevant Behaviour Detected	Datasets Referred to/Adopted
Probabilistic novelty detection for acoustic surveillance under real-world conditions [42]	2011	Applied Research	Audio	Clustering the GMMs of each sound sample for detecting outliers	Argument and fighting	Own dataset
Real-Time Abnormal Event Detection for Enhanced Security in Autonomous Shuttles Mobility Infrastructures [30]	2020	Applied Research	Video	Stacked Bidirectional LSTM Classifier, Spatiotemporal Autoencoder, and a Hybrid LSTM Classifier	Bag-snatching, pickpocketing, vandalism, and aggression	Data captured from Geneva Public Transport shuttles, NTU-RGB-D, and the UCSD
Real-world Anomaly Detection in Surveillance Videos [52]	2018	Applied Research	Video	MIL	Fighting, burglary, and robbery	UCF-Crime
Recent Advances in Video Analytics for Rail Network Surveillance for Security, Trespass and Suicide Prevention-A Survey [8]	2019	Survey	Video	N/A	N/A	CAVIAR
Scream and gunshot detection and localization for audio-surveillance systems [43]	2007	Applied Research	Audio	Two parallel GMM classifiers	Screams	Movie soundtracks, internet repositories, and own recorded
Sudden Event Recognition: A Survey [71]	2013	Survey	Video	N/A	N/A	Multi-camera Human Action Video, BEHAVE, CAVIAR

Table A1. Cont.

Title	Year	Research Type	Surveillance Type	Technology Adopted	Relevant Behaviour Detected	Datasets Referred to/Adopted
Suspicious Human Activity Recognition using 2D Pose Estimation and CNN [33]	2022	Applied Research	Video	CNN	Fighting and trespassing	UCI-HAR, WISDOM, RGB-D
Suspicious human activity recognition: a review [64]	2018	Review	Video	N/A	N/A	PETS 2006, PETS 2007, i-LIDS-abandoned baggage detection, VISOR, CVSG, CAVIAR, Bank, Fight CAVIAR, UCF-Crime
Two-Stream CNN Architecture for Anomalous Event Detection in Real World Scenarios [15]	2020	Applied Research	Video	Two-stream 2D-CNN	Abuse, arrest, burglary, fighting, robbery, stealing, and vandalism	UCF-Crime
Urban Anomaly Analytics: Description, Detection, and Prediction [28]	2022	Review	Video	N/A	N/A	Subway entrance, subway exit, UCSD, VIRAT, CUHK Avenue
Using Artificial Intelligence for Anomaly Detection Using Security Cameras [29]	2021	Applied Research	Video	CNN	Robbery	Data from the public transportation system of Grande Vitória, ES, Brazil
Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning [53]	2021	Applied Research	Video	Robust Temporal Feature Magnitude learning	Assault, burglary, robbery, shoplifting, stealing, vandalism, and fighting	ShanghaiTech Campus, UCF-Crime, XD-Violence, and UCSD

References

1. Stradling, S.; Carreno, M.; Rye, T.; Noble, A. Passenger perceptions and the ideal urban bus journey experience. *Trans. Policy* **2007**, *14*, 283–292. [\[CrossRef\]](#)
2. Quinones, L.M. Sexual harassment in public transport in Bogotá. *Trans. Res. Part A Policy Pract.* **2020**, *139*, 54–69. [\[CrossRef\]](#)
3. Lewis, S. *Sexual Harassment on the London Underground: Mobilities, Temporalities and Knowledges of Gendered Violence in Public Transport*; Loughborough University: Loughborough, UK, 2018.
4. Violence against Women an EU Wide Survey Main Results. European Union Agency for Fundamental Rights. 2014. Available online: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2014-vaw-survey-main-results-apr14_en.pdf (accessed on 3 May 2023).
5. Women's Safety and Security: A Public Transport Priority, 2018. International Transport Forum, OECD Publishing. Available online: https://www.itf-oecd.org/sites/default/files/docs/womens-safety-security_0.pdf (accessed on 3 May 2023).
6. Coppola, P.; Silvestri, F. Gender Inequality in Safety and Security Perceptions in Railway Stations. *Sustainability* **2021**, *13*, 4007. [\[CrossRef\]](#)
7. McAtamney, A.; Morgan, A. Key Issues in Antisocial Behaviour. Research in Practice. Australian Institute of Criminology. Available online: <https://www.aic.gov.au/publications/rip/rip5> (accessed on 29 January 2023).
8. Zhang, T.; Aftab, W.; Mihaylova, L.; Langran-Wheeler, C.; Rigby, S.; Fletcher, D.; Maddock, S.; Bosworth, G. Recent Advances in Video Analytics for Rail Network Surveillance for Security, Trespass and Suicide Prevention—A Survey. *Sensors* **2022**, *22*, 4324. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Anti-Social Behaviour, Crime Stoppers Western Australia. Available online: <https://www.crimestopperswa.com.au/keeping-safe/anti-social-behaviour/> (accessed on 19 January 2023).
10. Dezsö, C.L.; Ross, D.G. Does female representation in top management improve firm performance? A panel data investigation. *Strateg. Manag. J.* **2012**, *33*, 1072–1089. [\[CrossRef\]](#)
11. Woolley, A.W.; Chabris, C.F.; Pentland, A.; Hashmi, N.; Malone, T.W. Evidence for a collective intelligence factor in the performance of human groups. *Science* **2010**, *330*, 686–688. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Hunt, V.; Layton, D.; Prince, S. Diversity Matters. 2015. Available online: <https://www.mckinsey.com/business-functions/organization/our-insights/why-diversity-matters> (accessed on 29 January 2023).
13. Mohan, A.; Choksi, M.; Zaveri, M.A. Anomaly and Activity Recognition Using Machine Learning Approach for Video Based Surveillance. In Proceedings of the 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019; pp. 1–6.
14. Peng, H.-K.; Marculescu, R. Multi-Scale Compositionality: Identifying the Compositional Structures of Social Dynamics Using Deep Learning. *PLoS ONE* **2015**, *10*, e0118309. [\[CrossRef\]](#)
15. Majhi, S.; Dash, R.; Sa, P.K. Two-Stream CNN Architecture for Anomalous Event Detection in Real World Scenarios. In *Computer Vision and Image Processing-CVIP 2019*; Nain, N., Vipparthi, S., Raman, B., Eds.; Communications in Computer and Information Science; Springer: Singapore, 2019; Volume 1148.
16. Yao, H.; Hu, X. A survey of video violence detection. *Cyber-Phys. Syst.* **2021**, *9*, 1–24. [\[CrossRef\]](#)
17. Caetano, F.; Carvalho, P.; Cardoso, J. Deep Anomaly Detection for In-Vehicle Monitoring—An Application-Oriented Review. *Appl. Sci.* **2022**, *12*, 10011. [\[CrossRef\]](#)
18. Lee, S.; Kim, H.G.; Ro, Y.M. BMAN: Bidirectional Multi-Scale Aggregation Networks for Abnormal Event Detection. *IEEE Trans. Image Process.* **2020**, *29*, 2395–2408. [\[CrossRef\]](#)
19. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Peters, M.D.; Godfrey, C.M.; Khalil, H.; McInerney, P.; Parker, D.; Soares, C.B. Guidance for conducting systematic scoping reviews. *Int. J. Evid. Based Healthc.* **2015**, *13*, 141–146. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Arksey, H.; O'Malley, L. Scoping studies: Towards a methodological framework. *Int. J. Soc. Res. Methodol.* **2005**, *8*, 19–32. [\[CrossRef\]](#)
22. Levac, D.; Colquhoun, H.; O'Brien, K.K. Scoping studies: Advancing the methodology. *Implement. Sci.* **2010**, *5*, 1–9. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Colquhoun, H.L.; Levac, D.; O'Brien, K.K.; Straus, S.; Tricco, A.C.; Perrier, L.; Kastner, M.; Moher, D. Scoping reviews: Time for clarity in definition, methods, and reporting. *J. Clin. Epidemiol.* **2014**, *67*, 1291–1294. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Daudt, H.M.; van Mossel, C.; Scott, S.J. Enhancing the scoping study methodology: A large, inter-professional team's experience with Arksey and O'Malley's framework. *BMC Med. Res. Methodol.* **2013**, *13*, 1–48. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Basheer, A.M.I.; Halah, A.; Fatimah, A.; Dana, A.; Manar, A.; Munira, A.; Shoog, A.; Atta, R.; Mustafa, Y.; Gohar, Z. A Deep-Learning Approach to Driver Drowsiness Detection. *Safety* **2023**, *9*, 65.
26. Shebiah, R.N.; Arivazhagan, S. Ownership of abandoned object detection by integrating carried object recognition and context sensing. *Vis. Comput.* **2023**, *2023*, 1–26.
27. Babineau, J. Product review: Covidence (systematic review software). *J. Can. Health Libr. Assoc.* **2014**, *35*, 68–71. [\[CrossRef\]](#)

28. Zhang, M.; Li, T.; Yu, Y.; Li, Y.; Hui, P.; Zheng, Y. Urban Anomaly Analytics: Description, Detection, and Prediction. *IEEE Trans. Big Data* **2022**, *8*, 809–826. [\[CrossRef\]](#)
29. Affonso, G.A.; De Menezes, A.L.L.; Nunes, R.B.; Almonfrey, D. Using Artificial Intelligence for Anomaly Detection Using Security Cameras. In Proceedings of the 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Black River, Mauritius, 7–8 October 2021; pp. 1–5.
30. Tsiktsiris, D.; Dimitriou, N.; Lalas, A.; Dasygenis, M.; Votis, K.; Tzovaras, D. Real-Time Abnormal Event Detection for Enhanced Security in Autonomous Shuttles Mobility Infrastructures. *Sensors* **2020**, *20*, 4943. [\[CrossRef\]](#)
31. Ullah, W.; Ullah, A.; Haq, I.U.; Muhammad, K.; Sajjad, M.; Baik, S.W. CNN features with bi-directional LSTM for real-time anomaly detection in surveillance networks. *Multimed. Tools Appl.* **2021**, *80*, 16979–16995. [\[CrossRef\]](#)
32. Luo, W.; Liu, W.; Gao, S. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 341–349.
33. Dileep, A.S.; Nabilah, S.S.; Sreeju, S.; Farhana, K.; Surumy, S. Suspicious Human Activity Recognition using 2D Pose Estimation and Convolutional Neural Network. In Proceedings of the 2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 24–26 March 2022; pp. 19–23.
34. Nayak, R.; Pati, U.C.; Das, S.K. A comprehensive review on deep learning-based methods for video anomaly detection. *Image Vis. Comput.* **2021**, *106*, 104078. [\[CrossRef\]](#)
35. Hervás, M.; Fernandez-Medina, C.; Shiguihara-Juárez, P.; González-Valenzuela, R. Abnormal Behavior Detection: A Comparative Study of Machine Learning Algorithms Using Feature Extraction and a Fully Labeled dataset. In Proceedings of the 2019 International Conference on Information Systems and Computer Science (INCISCOS), Quito, Ecuador, 19–22 November 2019; pp. 62–67.
36. Ramachandran, S.; Palivela, L.H.; Vijayakumar, V.; Subramaniaswamy, V.; Abawajy, J.; Yang, L. An intelligent system to detect human suspicious activity using deep neural networks. *J. Intell. Fuzzy Syst.* **2019**, *36*, 4507–4518. [\[CrossRef\]](#)
37. Landi, F.; Snoek, C.G.M.; Cucchiara, R. Anomaly Locality in Video Surveillance. *arXiv* **2019**, arXiv:1901.10364.
38. Ullah, W.; Hussain, T.; Muhammad, K.; Heidari, A.A.; Ser, J.D.; Baik, S.W.; De Albuquerque, V.H. Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data. *Future Gener. Comput. Syst.* **2022**, *129*, 286–297. [\[CrossRef\]](#)
39. Liu, K.; Ma, H. Exploring Background-bias for Anomaly Detection in Surveillance Videos. In Proceedings of the 27th ACM International Conference on Multimedia, New York, NY, USA, 15 October 2019; pp. 1490–1499.
40. Ntalampiras, S.; Potamitis, I.; Fakotakis, N. An Adaptive Framework for Acoustic Monitoring of Potential Hazards. *EURASIP J. Audio Speech Music Process.* **2009**, *2019*, 594103. [\[CrossRef\]](#)
41. Clavel, C.; Vasilescu, I.; Devillers, L.; Richard, G.; Ehrette, T. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Commun.* **2008**, *50*, 487–503. [\[CrossRef\]](#)
42. Ntalampiras, S.; Potamitis, I.; Fakotakis, N. Probabilistic Novelty Detection for Acoustic Surveillance Under Real-World Conditions. *IEEE Trans. Multimed.* **2011**, *13*, 713–719. [\[CrossRef\]](#)
43. Valenzise, G.; Gerosa, L.; Tagliasacchi, M.; Antonacci, F.; Sarti, A. Scream and gunshot detection and localization for audio-surveillance systems. In Proceedings of the 2007 IEEE Conference on Advanced Video and Signal Based Surveillance, London, UK, 20 September 2007; pp. 21–26.
44. Yao, G.; Lei, T.; Zhong, J. A review of convolutional-neural-network-based action recognition. *Pattern Recognit. Lett.* **2019**, *118*, 14–22. [\[CrossRef\]](#)
45. al Zamil, M.G.H.; Samarah, S.; Rawashdeh, M.; Karime, A.; Hossain, M.S. Multimedia-oriented action recognition in smart city-based iot using multilayer perceptron. *Multimed. Tools Appl.* **2019**, *78*, 30315–30329. [\[CrossRef\]](#)
46. Xiao, S.; Wang, S.; Huang, Z.; Wang, Y.; Jiang, H. Two-stream transformer network for sensor-based human activity recognition. *Neurocomputing* **2022**, *512*, 253–268. [\[CrossRef\]](#)
47. Zhao, R.; Ali, H.; der Smagt, P.V. Two-stream rnn/cnn for action recognition in 3d videos. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 4260–4267.
48. Muhammad, K.; Ullah, A.; Imran, A.S.; Sajjad, M.; Kiran, M.S.; Sannino, G.; de Albuquerque, V.H.C. Human action recognition using attention based lstm network with dilated cnn features. *Future Gener. Comput. Syst.* **2021**, *125*, 820–830. [\[CrossRef\]](#)
49. Yenduri, S.; Perveen, N.; Chalavadi, V. Fine-grained action recognition using dynamic kernels. *Pattern Recognit.* **2022**, *122*, 108282. [\[CrossRef\]](#)
50. Blunsden, S.; Fisher, R.B. The BEHAVE Video Dataset: Ground Truthed Video for Multi-Person Behavior Classification. Available online: <https://homepages.inf.ed.ac.uk/rbf/PAPERS/unfbhavedata.pdf> (accessed on 20 November 2023).
51. Joo, H.K.; Vo, K.; Yamazaki, K.; Le, N. CLIP-TSA: CLIP-Assisted Temporal Self-Attention for Weakly-Supervised Video Anomaly Detection. *arXiv* **2022**, arXiv:2212.05136.
52. Sultani, W.; Chen, C.; Shah, M. Real-World Anomaly Detection in Surveillance Videos. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6479–6488.
53. Tian, Y.; Pang, G.; Chen, Y.; Singh, R.; Verjans, J.W.; Carneiro, G. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Online, 11–17 October 2021; pp. 4955–4966.

54. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Fangtao, S.; Wu, Z.; Yang, Z. Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision. In Proceedings of the European Conference on Computer Vision—ECCV 2020, Online, 23–28 August 2020; pp. 322–339.
55. Iantovics, L.B.; Enăchescu, C. Method for Data Quality Assessment of Synthetic Industrial Data. *Sensor* **2022**, *22*, 1608. [CrossRef] [PubMed]
56. Courty, N.; Allain, P.; Creusot, C.; Corpetti, T. Using the Agoraset dataset: Assessing for the quality of crowd video analysis methods. *Pattern Recognit. Lett.* **2014**, *44*, 161–170. [CrossRef]
57. Kawamura, A.; Yoshimitsu, Y.; Kajitani, K.; Naito, T.; Fujimura, K.; Kamijo, S. Smart camera network system for use in railway stations. In Proceedings of the 2011 IEEE International Conference on Systems, Man, and Cybernetics, Maui, HI, USA, 9–12 October 2011; pp. 85–90.
58. Ren, J.; Xia, F.; Liu, Y.; Lee, I. Deep Video Anomaly Detection: Opportunities and Challenges. In Proceedings of the 2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand, 7–10 December 2021; pp. 959–966.
59. Pawar, K.; Attar, V. Deep learning approaches for video-based anomalous activity detection. *World Wide Web* **2019**, *22*, 571–601. [CrossRef]
60. Wang, T.; Meina Qiao, M.; Deng, Y.; Zhou, Y.; Wang, H.; Lyu, Q.; Snoussi, H. Abnormal event detection based on analysis of movement information of video sequence. *Optik* **2018**, *152*, 50–60. [CrossRef]
61. Santhosh, K.K.; Dogra, D.P.; Roy, P.P. Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey. *ACM Comput. Surv.* **2021**, *53*, 119–126. [CrossRef]
62. Vu, T.-H.; Boonaert, J.; Ambellouis, S.; Taleb-Ahmed, A. Multi-Channel Generative Framework and Supervised Learning for Anomaly Detection in Surveillance Videos. *Sensors* **2021**, *21*, 3179. [CrossRef]
63. Ferryman, J. Video Surveillance Standardisation Activities, Process and Roadmap: ERNCIP Thematic Group Video Surveillance for Security of Critical Infrastructure. Technical Report JRC103650; Joint Research Centre (JRC). 2016. Available online: <https://data.europa.eu/doi/10.2788/92267> (accessed on 22 January 2023).
64. Tripathi, R.K.; Jalal, A.S.; Agrawal, S.C. Suspicious human activity recognition: A review. *Artif. Intell. Rev.* **2018**, *50*, 283–339. [CrossRef]
65. Zhu, Y.; Newsam, S. Motion-Aware Feature for Improved Video Anomaly Detection. *arXiv* **2019**, arXiv:1907.10211.
66. Ramzan, M.; Abid, A.; Khan, H.U.; Awan, S.M.; Ismail, A.; Ahmed, M.; Ilyas, M.; Mahmood, A. A Review on State-of-the-Art Violence Detection Techniques. *IEEE Access* **2019**, *7*, 107560–107575. [CrossRef]
67. Ye, M.; Peng, X.; Gan, W.; Wu, W.; Qiao, Y. AnoPCN: Video Anomaly Detection via Deep Predictive Coding Network. In Proceedings of the 27th ACM International Conference on Multimedia, New York, NY, USA, 27 October 2019; pp. 1805–1813.
68. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future Frame Prediction for Anomaly Detection—A New Baseline. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6536–6545.
69. Markovitz, A.; Sharir, G.; Friedman, I.; Zelnik-Manor, L.; Avidan, S. Graph Embedded Pose Clustering for Anomaly Detection. *arXiv* **2022**, arXiv:1912.11850.
70. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; Van den Hengel, A. Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 11–17 October 2019; pp. 1705–1714.
71. Suriani, N.S.; Hussain, A.; Zulkifley, M.A. Sudden Event Recognition: A Survey. *Sensors* **2013**, *13*, 9966–9998. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.