applied sciences

MDPI

Review

Deep Anomaly Detection for In-Vehicle Monitoring—An Application-Oriented Review

Francisco Caetano [1,2], Pedro Carvalho [1,3] and Jaime Cardoso [1,2,*]

[1] INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
[2] Faculty of Engineering (FEUP), University of Porto, 4200-465 Porto, Portugal
[3] School of Engineering (ISEP), Polytechnic of Porto, 4200-072 Porto, Portugal
* Correspondence: jaime.cardoso@inesctec.pt

Abstract: Anomaly detection has been an active research area for decades, with high application potential. Recent work has explored deep learning approaches to the detection of abnormal behaviour and abandoned objects in outdoor video surveillance scenarios. The extension of this recent work to in-vehicle monitoring using solely visual data represents a relevant research opportunity that has been overlooked in the accessible literature. With the increasing importance of public and shared transportation for urban mobility, it becomes imperative to provide autonomous intelligent systems capable of detecting abnormal behaviour that threatens passenger safety. To investigate the applicability of current works to this scenario, a recapitulation of relevant state-of-the-art techniques and resources is presented, including available datasets for their training and benchmarking. The lack of public datasets dedicated to in-vehicle monitoring is addressed alongside other issues not considered in previous works, such as moving backgrounds and frequent illumination changes. Despite its relevance, similar surveys and reviews have disregarded this scenario and its specificities. This work initiates an important discussion on application-oriented issues, proposing solutions to be followed in future works, particularly synthetic data augmentation to achieve representative instances with the low amount of available sequences.

Keywords: anomaly detection; deep learning; computer vision; anomaly locality; in-vehicle monitoring

Citation: Caetano, F.; Carvalho, P.; Cardoso, J. Deep Anomaly Detection for In-Vehicle Monitoring—An Application-Oriented Review. *Appl. Sci.* **2022**, *12*, 10011. https://doi.org/10.3390/app121910011

Academic Editor: Andrea Prati

Received: 31 August 2022
Accepted: 30 September 2022
Published: 5 October 2022

1. Introduction

The proliferation of cameras and the growing availability of cheap storage, coupled with the increasing demand for security, have fuelled the development of ever more complex video surveillance systems. While the task of capturing the images of possible transgressions has been greatly facilitated, appointing human controllers (e.g., a security guard) with the duty of analysing repetitive and monotonous images, as well as a multiple-camera perspective, represents a critical flaw. Such an exhausting human effort makes it very difficult for the controller to remain vigilant at all times, which might lead to abnormal events going unnoticed. In safety-critical domains, such as the detection of suspect packages in airports or train stations, neglected occurrences may have dangerous results.

The identification of unexpected events, behaviours or objects can be recast as an anomaly detection problem [1,2]. The application of deep anomaly detection methods is essential to develop new surveillance and monitoring systems that do not rely solely on human supervision, reducing the risk of the aforementioned drawbacks. Despite years of research and development, the detection of anomalies in videos remains challenging, and it differs from the traditional classification problem in two large aspects. Firstly, new kinds of anomalies are constantly arising, making it virtually impossible to list all of them. Secondly, the task of collecting sufficient negative samples is costly due to their rarity. A popular method for deep anomaly detection consists of using videos of normal events as training data. A test set is then used to detect the abnormal events which would not conform to the

model that was trained [3,4]. This approach aims to circumvent the difficulty of gathering sufficient samples representing anomalies. Most of the frames that these systems analyse represent normal scenarios; hence, the gathered data are representative of this low ratio of abnormal snippets, and fully supervised approaches are not viable.

Deep learning approaches to the detection of visual data instances that markedly digress from regular sequences have been mostly focusing on outdoor video-surveillance scenarios, mainly regarding abnormal behaviour and suspicious or abandoned object detection. A pertinent research opportunity for anomaly detection that has been overlooked in the accessible literature is posed by in-vehicle monitoring specially using solely visual data. With the increasing relevance of public transport in urban mobility, several funded projects aiming to develop autonomous surveillance systems in this area have appeared. For instance, Prevent PCP involves some of the biggest transport operators in Europe, which consider that a concerted effort is required to develop systems capable of detecting abnormal behaviours that put passengers' safety at risk. In the initial stages of this project, the lack of task-oriented datasets has been noted; an effort to acquire and label the required footage to build a dedicated dataset has been programmed. However, in-vehicle monitoring is not limited to public transport, in which large crowds must be monitored; the advent of Shared Autonomous Vehicles [5], which do not have a driver responsible for maintaining the well-being of passengers, must be accompanied by competent and reliable autonomous in-vehicle surveillance systems.

The development of robust solutions for in-vehicle monitoring is not straightforward, as the conditions in which it must operate are very challenging and different from those that the methods that cover outdoor video surveillance face. Nonetheless, a recapitulation of relevant state-of-the-art techniques and available resources is essential to investigate their applicability and to achieve a deeper understanding of the potential issues raised by the new scenario that these methods did not contemplate. Furthermore, the same principle must be applied to the analysis of available datasets to train and benchmark such models. It is essential to acknowledge if any portion of the available datasets is representative of the real-world settings that the systems will face; if not, the possibility of repurposing and adapting these instances should be studied. The current challenges of developing an application-oriented solution to in-vehicle monitoring have two distinct origins. On the one hand, there are potential issues that are directly linked to the characteristics of the application, which are specifically manifested by the absence of public datasets explicitly dedicated to in-vehicle monitoring. Additionally, the importance of actor independence in Shared Autonomous Vehicles, moving backgrounds and frequent illumination changes caused by the movement of the vehicle are important factors to consider. On the other hand, there are current limitations that are transversal to every anomaly detection technique that has been proposed. As Pang et al. [6] denote, a series of complex detection challenges remain largely unsolved and are yet to be fully addressed by deep anomaly detection. The first of these challenges is the low anomaly detection recall rate caused by their rare and heterogeneous nature; as they are difficult to identify, sophisticated anomalies are missed. Additionally, since the candidate pool of anomalies is often unbounded, the strategies that these methods employ to deal with novelty must not be overlooked.

As a consequence of the previously mentioned difficulty of gathering abnormal samples for training or validation, there is a significant effort to achieve high data efficiency for learning normality and abnormality. Fully supervised anomaly detection is, for the time being, a virtually impossible endeavour, mainly due to the high cost of collecting large-scale data or generating sufficiently broad artificial dataset solutions. When some labels for anomaly classes are available, they might be incomplete, inexact (e.g., coarse-grained) or inaccurate. The subject of actor independence is relevant as well, as the ones present in the training data could generate a bias due to their lack of representativity (e.g., height, gender, age, type of clothes). In addition to learning expressive representations with a small amount of data, it is also essential to learn models that are generalisable to novel anomalies. This theme also extends to noise-resilient anomaly detection, with noise being equivalent to

mislabelled data or unlabelled anomalies. The amount of noise not only differs significantly from dataset to dataset, but it is also irregularly distributed in the data space. Noise-resilient models can leverage this incomplete data to achieve better performance and robustness.

Most currently developed methods are committed to detecting individual instances that are anomalous, which are often regarded as point anomalies. However, more complex anomalies, such as conditional and group anomalies, comprise objectively different dynamics and behaviours. Conditional anomalies also refer to individual anomalous instances, but they only represent abnormal behaviour when they occur in a specific context. Group anomalies are anomalous as a whole, although the isolated behaviour of every member might not be abnormal. Furthermore, many applications require the detection of anomalies with multiple data sources, heterogeneous (e.g., video and audio) or not (e.g., multiple surveillance cameras). The complexity of these systems is yet to be properly addressed by deep anomaly detection strategies, even though high-dimensional anomaly detection has been a long-standing problem [7]. Identifying intricate feature interactions and couplings is already a challenge when temporal and spatial interdependency relationships are considered.

The success of Machine Learning (ML) has led to a growing interest in the development of Artificial Intelligence (AI) applications capable of providing explanations to their decisions, which are often called Explainable AI [8]. This information is essential for users to trust, understand and manage these applications. However, every explanation is set within a context that depends on what is expected of the AI system. For in-vehicle monitoring, anomaly identification should be one of the main points of interest, and it should generally be paired with anomaly classification. The detected anomalies should be coupled with cues that demonstrate why a specific data instance is abnormal. The simplest implementation of this practice is to spatially identify the anomaly in a frame (e.g., with a bounding box or a GradCAM activation map). However, most anomaly detection studies focus on detection performance only, ignoring the capability of illustrating the identified anomalies. The complexity of the anomalies calls for developing visually interpretable anomaly detection models, as the provided cues could be essential to identify problems such as under-represented groups. That said, this work appears as the first aggregated critical review on the applicability of deep video anomaly detection to in-vehicle monitoring, making the following three major contributions:

- Review of a large number of state-of-the-art methods for deep video anomaly detection, aiming to explain their framework and implementation, thus providing a deeper understanding of potential issues raised by the new scenario that these methods did not contemplate. Benchmarks for their performance were compiled as well as publicly accessible source codes to evaluate the ease of applicability;
- Review of a large number of datasets with real anomalies that are used to benchmark state-of-the-art models, investigating if any portion of the available datasets is representative of the real-world settings for in-vehicle monitoring or if the sequences can be repurposed for this matter. As public datasets dedicated to in-vehicle monitoring are lacking, this analysis is vital;
- This work initiates an important discussion on application-oriented issues related to deep anomaly detection for in-vehicle monitoring. Other surveys and reviews have disregarded this scenario and its specificities, despite its relevance, as shown by the listed funded projects that seek application-oriented solutions for in-vehicle monitoring. Possible solutions were proposed, aiming to follow up on future work.

This document is organised as follows: Section 2 presents the working principles, different approaches, and state-of-the-art works in deep anomaly detection for video sequences. Section 3 reviews currently used datasets to train and benchmark models for anomaly detection. Moreover, Section 4 discusses funded projects that seek the exploration of in-vehicle monitoring, listing available resources that serve as a starting point for developing application-oriented solutions. Section 5 examines the challenges that this new

scenario of application faces as well as the opportunities that arise with its exploration. Finally, Section 6 presents the main conclusions of this review.

2. Literature Review on Deep Anomaly Detection

Anomaly detection in videos is commonly framed as a one-class classification task. Such a strategy requires a training set strictly composed of normal events, whilst the videos for testing represent normal and abnormal events. This approach can be defined as a semi-supervised strategy, which is a term regularly used in the categorisation of deep anomaly detection approaches [9]. Furthermore, weakly supervised techniques have also been referred to in the literature as a category for such approaches [10]. They differ from semi-supervised strategies by obtaining video-level labels, which allow for the training of the models using normal and abnormal snippets. As far as fully supervised strategies are concerned, the cost of collecting and labelling large-scale data for this purpose is unbearable. However, the inclusion of synthetic data and a focus on locality have been studied as assets to leverage the advantages of fully supervised approaches, introducing them to semi-supervised and weakly supervised strategies.

2.1. Evaluation Metrics

In the literature on anomaly detection [11,12], a prominent evaluation metric is the Receiver Operation Characteristic (ROC), which is obtained by gradually changing the threshold of the regularity score. The regularity score is used to judge whether the input frame is normal or abnormal by manually defining a threshold. The optimal value of this parameter is relevant, since a higher threshold leads to a higher false negative rate, while a lower one leads to a higher false negative rate. Then, the Area Under the Curve (AUC) is cumulated to a scalar for performance evaluation with a higher value indicating better performance. As AUC is the only metric that is present in every referenced work, it was leveraged for performance evaluation in this paper. To achieve a more detailed understanding of a certain model, precision, recall, true positive, and false alarm should also be considered. Although these metrics are not as popular as Frame AUC, their inclusion provides interesting indicators to estimate the potential success of a real-world application of a certain model (e.g., knowing if it is prone to false alarms).

2.2. Semi-Supervised Strategies

Generally, semi-supervised methods for anomaly detection in the literature fall in the category of One-Class Classification (OCC). In practice, it is quite frequent that normal events have a good representation, as they represent the majority of the captured sequences, whilst abnormal cases are rare, and the abnormal class is ill-defined. In these cases, the abnormality is detected based on the information learnt from the normal class only. One-Class Classification is present in reconstruction-based and prediction-based semi-supervised methods that employ the same basic principle; both try to generate the entirety or patches of a frame, evaluating their similarity to the ground truth. These models are trained on normality and assume that it is not possible to properly reconstruct an abnormal event that has never been learnt. Hence, a frame that greatly differs from the captured one is likely to represent abnormal or unexpected events. The main difference between both types of semi-supervised methods regards temporal information, as illustrated in Figure 1. On the one hand, reconstruction-based methods try to reconstruct the current frame, using previous and present information. On the other hand, prediction-based methods use the previous frames to compute a prediction of the following one. Recently, semi-supervised strategies have been increasingly focusing on video frame prediction due to its potential applications in unsupervised video representation learning. However, existing methods of this type deliver suboptimal results, especially when compared to newer methods that use weakly supervised techniques, due to their insufficient modelling of temporal information. Moreover, they suffer from inefficient training for implementing adversarial techniques or additional losses [13]. The lack of prior knowledge of abnormality is usually a cause of

overfitting of the training data, not enabling a proper way to distinguish abnormal from normal events [6].

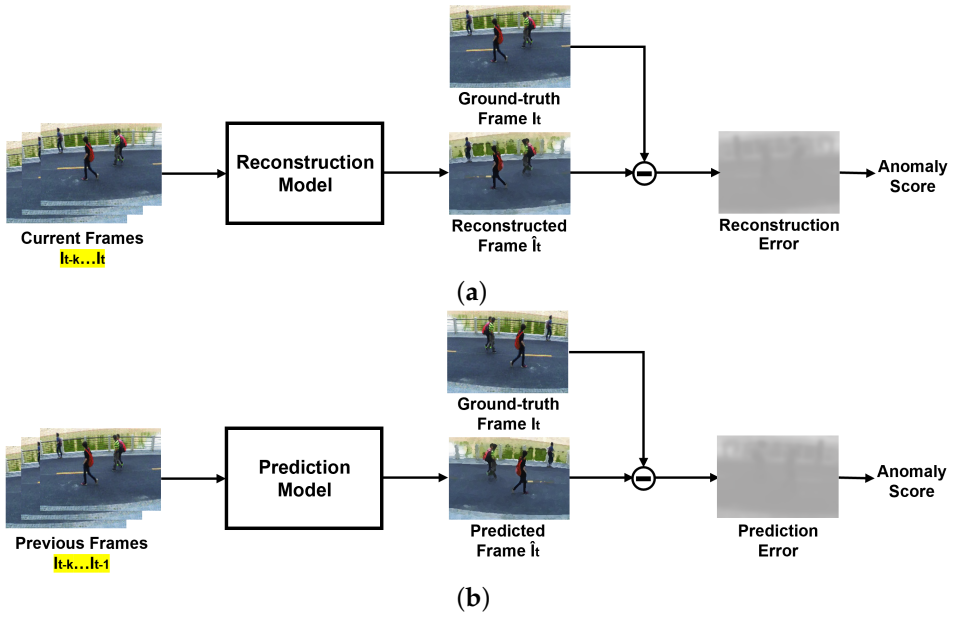

Figure 1. Pipelines of semi-supervised methods for video anomaly detection. They differ on the use of the current frame for generation, as highlighted in yellow. (**a**) Generic pipeline of reconstruction methods. (**b**) Generic pipeline of prediction methods.

2.2.1. Reconstruction-Based Methods

Reconstruction-based models were amongst the first deep learning approaches to use anomaly detection in video instances and were built on the assumption that a model trained on normal events cannot properly reconstruct an abnormal event that it has never seen. A precursor of this work was the approach of Xu et al. [14], which relied on stacked denoising autoencoders to learn appearance and motion features that were posteriorly fed to multiple one-class Support Vector Machine (SVM) models.

The deep autoencoder, ConvAE, proposed by Hasan et al. [15] diverged from using autoencoders simply as feature extractors, becoming the first anomaly detection approach to leverage the reconstruction error as an estimator for abnormality. Although the model takes multiple frames as input, temporal information is lost, since the convolution operations are performed spatially. Therefore, this work was quickly followed by Conv3D-AE [16,17], suggesting a 3D convolutional neural network to encode the motion and content information of a sequence of frames, using a deconvolutional network to reconstruct those frames. However, 3D convolution has proved to be unable to properly encode motion [18,19]. Additionally, these methods predict the anomaly for all clips in a video, making it time consuming to obtain a frame-level prediction.

A Convolutional Neural Network (CNN) and ConvLSTM were integrated with an autoencoder in ConvLSTM-AE [20] to learn the regularity of appearance and motion for ordinary moments. Although LSTMs and Recurrent Neural Networks (RNNs) are effective for sequential data processing and succesfuly encode the motion in videos [21], they are hard to be interpreted; hence, several works focused on adapting sparse coding techniques and interpretable RNNs to anomaly detection [4,22].

As Liu et al. [3] denoted, autoencoder-based approaches are at times able to accurately reconstruct abnormal frames based on the provided inputs, leading to missing their detection. This undermines the assumption that the reconstruction error of an abnormal frame is significantly different than that of a normal one. To deal with this drawback, a memory module was added to the autoencoder by Gong et al. [23], creating MemAE, a memory-augmented autoencoder. An input was firstly encoded, and then, the com-

pressed information was queried to retrieve the relevant memory items for reconstruction. During testing, the memory module was fixed, and the reconstruction used the stored data.

Several GAN-based approaches were also proposed. For instance, Ravanbakhsh et al. [24] trained two conditional GANs on normal frames and corresponding optical-flow images to comprehend the correct representation of the scene normality. Furthermore, Ganokratanaa et al. [25] proposed an attempt to improve anomaly localisation at the pixel level by introducing a technique to localise pixels that belong to the anomalous objects, which is called Edge Wrapping. Finally, Ye et al. [26] tried to reduce the gap between the two semi-supervised methods by unifying reconstruction and prediction methods in an end-to-end framework.

2.2.2. Prediction-Based Methods

As far as prediction-based approaches are concerned, these aim to predict future frames based on an input consisting of previous frames. This method was introduced by Liu et al. [3] and assumes that normal events are predictable, while abnormal ones are not. Future Frame Prediction [3] proposed strategies to impose consistency on the generated images by applying intensity and gradient constraints. The former assures the similarity of all pixels in the RGB space, and the latter sharpens the generated images.

Taking inspiration from the *cloze test* used in language understanding, Yu et al. [27] proposed the prediction of erased patches of incomplete video events, fully exploiting temporal information in the video. Nonetheless, it still depended on pixel-wise constraints to regularise the prediction task, ignoring the correlation between optical flow and the frames. Unlike these approaches, Chen et al. [28] aimed to explore the information contained in the anterior and posterior snippets of a given frame within a video. For that purpose, it modelled the relationship between appearance and motion through a multi-modal discriminator (MD). The MD was based on the work of Radford et al. [29] and constructs an association between appearance and motion predictions; the discriminator was fed the concatenation of an erased patch and its motion to learn to classify fake and real pairs. The temporal relationships in the video sequence were also considered.

Georgescu et al. [30] proposed some alterations to middle-frame prediction [31], innovating by learning the discrimination of moving objects, which is referred to as the arrow of time. Additionally, it studied motion irregularity prediction and model distillation, the latter being an adaptation of Bergmann et al. [32]. Essentially, model distillation considers both classification and detection information, producing large prediction discrepancies when anomalies occur. This approach was inspired by the object-centric perspective of Ionescu et al. [33], which employed an object detector on each frame, applying a convolutional autoencoder to learn deep unsupervised representations for a one-versus-rest classification. Several works continued the exploration of this research [27,34,35]. However, Georgescu et al. [30] only retained the object detector from these approaches, focusing its analysis on the detected objects.

The main drawbacks of semi-supervised approaches are the lack of consideration for the diversity of normal patterns and the ability of deep learning techniques to correctly recreate abnormal video frames based on already abnormal inputs. To this end, Park et al. [36] proposed a memory module that updates items in the memory while assuring that these represent prototypical patterns of normal data. Similarly, Cai et al. [37] attempted to assure appearance and motion consistency through modality memory pools. Two separate pools were created to store this information: one comprising appearance features and the other consisting of the motion features, guaranteeing a robust feature representation of normality.

Not every approach focuses specifically on the error of the generated frame, one example being suggested by Ramachandra et al. [38]. The authors employed a Siamese network capable of learning a metric between spatiotemporal video patches. The dissimilarity between patches was used to estimate the level of abnormality of a frame. The work of Lee et al. [31] explored multi-level frameworks, generating inter-frame predictions and an attention map, which were fed to an appearance–motion joint detector to evaluate the

normality score. The performance of semi-supervised methods can be compared through the benchmark results shown in Table 1. The datasets referenced in the Table are fully explored in Section 3.

Table 1. Frame-level AUC scores of several semi-supervised methods, benchmarked on CUHK Avenue [11] and ShanghaiTech Campus [4] datasets. Results compiled by Feng et al. [39] Legend: †—Computed by Luo et al. [4].

Year	Method	AUC Score (%)	
		CUHK Avenue	ShanghaiTech
2016	Conv-AE [15]	70.2	60.85 †
2017	ConvLSTM-AE [20]	77.0	-
	S-RNN [4]	81.7	68.0
2018	FFP [3]	85.1	72.8
2019	Mem-AE [23]	83.3	71.2
	Object-Centric [33]	90.4	84.9
2020	MNAD [36]	88.5	70.5
	VEC [27]	90.2	74.8
2021	SSMT [30]	86.9	83.5
	ROADMAP [13]	88.3	76.6
	AMMC [37]	86.6	73.7
2022	BDPN [28]	90.3	78.1

2.3. Weakly Supervised Strategies

Weakly supervised video anomaly detection strategies are amongst the best-performing approaches for this task. At the expense of an additional low-intensity annotation effort, a better anomaly classification accuracy can be achieved. Essentially, weakly supervised approaches can be subdivided into two classes, encoder-agnostic and encoder-based methods. Encoder-agnostic methods [40–42] leveraged task agnostic features of videos extracted from a vanilla feature encoder (e.g., I3D [43]) to estimate the anomaly scores of each frame. In these methods, only the classifier was trained. On the other hand, encoder-based methods [44,45] trained both the feature encoder and classifier simultaneously.

Weakly supervised strategies are considered to be a feasible method due to their competitive performance. Sultani et al. [40] introduced the use of video-level labels in the tasks of anomaly detection in videos by presenting UCF-Crime, which is a large-scale video dataset for training and testing weakly supervised anomaly detection approaches. Along with this strategy, Sultani et al. [40] proposed a deep Multiple Instance Learning (MIL) ranking framework to detect anomalies, as illustrated in Figure 2. In essence, MIL takes a video as a bag and clips the video as separate instances. The bag generated from an abnormal instance is called a positive bag, and it must contain at least one abnormal snippet. The negative bag, generated from normal videos, contains no abnormal snippets. The instance-level anomaly scores are learnt through the bag-level labels. Several papers followed the MIL framework, suggesting improvements to the method. The inner-bag score gap regularisation was introduced by Zhang et al. [41] to increase the gap between the lowest and highest scores in a positive bag and reduce it in a negative one. Wan et al. [42] proposed a dynamic MIL-loss and centre-guided regularisation; the former enlarged the interclass dispersion, and the latter reduced the intraclass distance of normal snippets. Additionally, Zhu et al. [44], in an encoder-based approach, suggested an attention-based MIL model capable of encoding motion-aware features by using an autoencoder based on optical flow. To unify the representation learning and anomaly score learning, a temporal feature ranking loss was presented by Tian et al. [46]. This approached proved to be capable of achieving a better separation of normal and abnormal features, improving the exploration of the weak-labelling strategy in comparison to the previous MIL methods.

Zhong et al. [45] denoted that the methods that used Multiple Instance Learning suffered from error propagation throughout the training. If the model incorrectly predicted anomalous instances in the positive bag, the error would affect subsequent instance selection. To tackle this problem, Zhong et al. [45] reformulated the task as a binary classification under a noisy label problem and suggested the use of a Graph Convolution Neural (GCN) network to correct low-confidence anomaly scores, replacing them with high-confidence ones, i.e., clear the label noise. In the GCN, two characteristics of a video were considered to correct the label noise: temporal consistency, and feature similarity. Feature similarity assures that the abnormal snippets share similar characteristics; temporal consistency guarantees that abnormal instances appear in temporal proximity of each other. Even though this work achieved better accuracy in the identification of anomalies when compared to MIL-based approaches, training both a GCN and MIL is computationally expensive and may cause unstable performance due to unconstrained latent space.

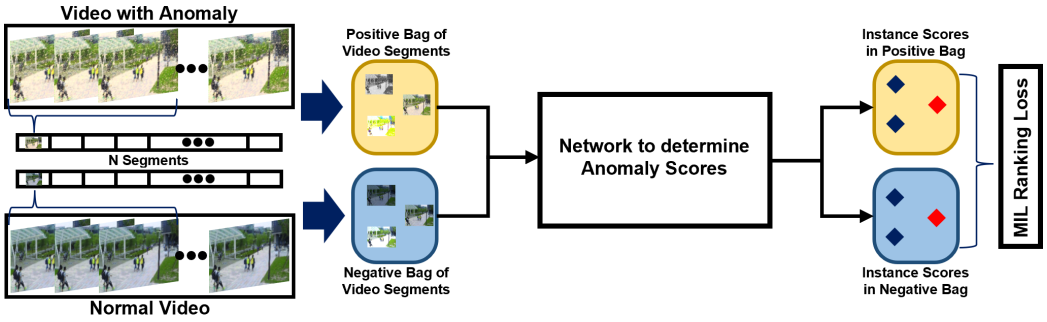

Figure 2. Architecture of a typical MIL-based method, such as the one proposed by Sultani et al. [40].

Li et al. [47] proposed another approach capable of addressing the shortcomings of MIL-based methods. The authors of this paper used a Multi-Sequence Learning (MSL) method, opting for choosing the sequence with the highest sum of anomaly scores instead of the instance with the highest score, reducing the probability of incorrect selection. This method encoded the extracted snippet features via a multi-layer Convolutional Transformer Encoder, using VideoSwin [48] as the backbone to extract them. The benchmark results shown in Table 2 indicate that VideoSwin performed consistently better than traditional feature extractors, such as C3D-RGB and I3D-RGB.

The goal of video anomaly detection must be the prediction of fine-grained anomaly scores. Although some works focused on detecting anomalies in an offline or coarse-grained manner, these do not allow real-time monitoring, reducing the interest in applying them to real-world scenarios. Instead of using video-level labels as pseudo-labels, Feng et al. [49] suggested the use of the learnt pseudo-labels to optimise the feature encoder. This method was capable of working in an online fine-grained manner. In a similar fashion to Li et al. [47], a two-stage self-training strategy was used.

Table 2. Frame level AUC scores of several weakly supervised methods, benchmarked on ShangaiTech Campus [4] and UCF-Crime [40] datasets. Results compiled by Feng et al. [39] Legend: †—Computed by Wan et al. [42].

Year	Method	Feature Extractor	AUC Score (%)	
			ShanghaiTech	UCF-Crime
2018	Sultani et al. [40]	C3D-RGB	86.3 †	75.41
2019	IBL [41]	C3D-RGB	-	78.66
	GCN-Anomaly [45]	C3D-RGB	76.44	81.08
		TSN-Flow	84.13	78.08
		TSN-RGB	84.44	82.12
	Motion-Aware [44]	PWC-Flow	-	79.0

Table 2. *Cont.*

Year	Method	Feature Extractor	AUC Score (%)	
			ShanghaiTech	UCF-Crime
2020	AR-Net [42]	I3D-RGB & I3D Flow	91.24	-
2021	MIST [49]	I3D-RGB	94.83	82.30
	RTFM [46]	I3D-RGB	97.21	84.30
	CRFD [50]	I3D-RGB	97.48	84.89
	BD-LSTM [51]	ResNet-50	-	85.53
2022	MSL [47]	C3D-RGB	94.81	82.85
		I3D-RGB	96.08	85.30
		VideoSwin-RGB	97.32	85.62

Ullah et al. [51] developed an approach that aims to reduce the processing time required for deep anomaly detection. For this purpose, the features extracted from the sequence of frames were fed to a Bi-directional Long Short-term Memory (BD-LSTM) model, which differs from a regular LSTM by depending not only on the previous frames but also on the upcoming ones. This work was followed by Ullah et al. [52], decomposing the anomaly detection process into two stages. Firstly, a Raspberry Pi, a resource-limited device, runs a lightweight version of the network. If an anomaly is detected, the captured data are transmitted to a cloud centre for a detailed analysis. The performance of weakly supervised methods can be evaluated through the benchmark results illustrated in Table 2. The datasets referenced in the table are fully explored in Section 3.

2.4. Fully Supervised Strategies

The exploration of fully supervised anomaly detection remains limited by the high cost of collecting large-scale data or generating sufficiently broad artificial dataset solutions. However, semi-supervised and weakly supervised methods, despite their recent improvements, have not fully addressed their limitations, such as background bias. Liu et al. [53] conducted a series of experiments to validate the existence of a background-bias phenomenon, i.e., the tendency for deep neural networks to learn the background information rather than the anomaly pattern. To tackle this, a portion of UCF-Crime [40] was re-annotated with temporal and spatial labels, the latter being represented by bounding boxes. This new information was used to feed an end-to-end framework with a designed region loss, explicitly guiding the model to focus on the anomalous region.

A similar approach was implemented by Landi et al. [54], focusing on spatiotemporal tubes instead of the entirety of video segments containing full frames. UCFCrime2Local, an enriched subsection of 100 burglary and assault sequences from UCF-Crime [40], was presented as a separate dataset for anomaly detection with bounding box supervision in its train and test set. Furthermore, the proposed trainable model for anomaly detection was designed to be capable of dealing with different abnormal locations in the same video segment, and the obtained results demonstrate that locality is robust to different kinds of errors in the tube extraction phase at test time. Moreover, the proposed model was able to provide spatiotemporal proposals for unseen surveillance videos leveraging only video-level labels, enlarging the anomaly dataset without additional human labelling.

Acsintoae et al. [55] presented UBnormal, a synthetic dataset generated in Cinema4D that introduces abnormal events annotated at the pixel level in the training set, enabling the use of fully supervised anomaly detection methods for the first time. Nevertheless, simulated scenes belong to a different data distribution than natural scenes. Hence, the behaviour of fully supervised models trained on this dataset might be unclear when it is applied to real-world scenes. To bridge the gap between synthetic and real-world datasets, Acsintoae et al. [55] proposed the translation of simulated objects from UBnormal to datasets such as Avenue [11] and ShanghaiTech [4] using a CycleGAN [56]. The results presented in the paper demonstrate that this hybrid solution of data augmentation can

enhance the performance of state-of-the-art anomaly detection models. The performance of the aforementioned methods can be evaluated through the benchmark results exhibited in Table 3. Furthermore, a summary of the most relevant surveyed methods with different supervision strategies is provided in Table 4. The datasets referenced in both tables are fully explored in Section 3.

Table 3. Frame level AUC scores of several fully supervised methods, benchmarked on CUHK Avenue [11], ShanghaiTech Campus [4] and UCF-Crime [40] datasets. Results compiled by Feng et al. [39]. Legend: †—The subset UCFCrime2Local [54] was used.

Year	Method	AUC Score (%)		
		UCF-Crime	CUHK Avenue	ShanghaiTech
2019	Liu et al. [53]	82.0	-	-
	Landi et al. [54]	77.52 †	-	-
2022	UBnormal [55]	-	93.2	83.7

Table 4. Summary of the most relevant surveyed methods. Abbreviations: A = CUHK Avenue [11], SU = Subway (Entry and Exit) [57], P = UCSD Pedestrian [12], U = UMN [58], S = ShanghaiTech Campus [4], UC = UCF-Crime [40], X = XD-Violence [59], UC2L = UCFCrime2Local [54].

Type	Year	Method	Datasets	Major Contributions
Semi-supervised	2016	Conv-AE [15]	A, SU, P	Estimated abnormality through the reconstruction error of the learnt AE.
	2017	ConvLSTM-AE [20]	A, SU, P	A CNN for appearance encoding and a ConvLSTM for memorising motion information of past frames were integrated with the AE.
		S-RNN [4]	A, SU, P, S	Temporally coherent Sparse Coding mapped to a Stacked-RNN, improving parameter optimisation and anomaly prediction speed.
	2018	FFP [3]	A, P, S	Future frames were predicted with motion and intensity constraints and were compared with the ground truth to detect anomalies.
	2019	Mem-AE [23]	A, P, S	Given an input, it used the encoded information as a query to retrieve the most relevant memory items for reconstruction.
		Object-Centric [33]	P, U, S	Object-centric AE to encode motion and appearance information, paired with a one-versus-rest classifier to separate normality clusters.
		BMAN [31]	A, P, U, S	Introduced an inter-frame predictor to encode normal patterns, which is used to detect abnormal events in an appearance-motion joint detector.
	2020	VEC [27]	A, P, S	Prediction of erased patches of incomplete video events, fully exploiting temporal information in the video.
		MNAD [36]	A, P, S	Added a memory module to record prototypical patterns of normal data in memory items, training it with compactness and separateness losses.
	2021	SSMT [30]	A, P, S	Considered the discrimination of moving objects and objects in consecutive frames; reconstruction of object-specific appearance information.
		AMMC [37]	A, P, S	Combined appearance and motion features to obtain an essential and robust representation of regularity.
		ROADMAP [13]	A, P, S	Used a frame prediction network that handles objects and different scales better; introduced a noise tolerance loss to mitigate background noise.
	2022	BDPN [28]	A, P, S	Introduced three constraints to regularise the prediction task from pixel-wise, cross-modal, and temporal-sequence levels.

Table 4. *Cont.*

Type	Year	Method	Datasets	Major Contributions
Weakly-supervised	2018	Sultani et al. [40]	UC	Learnt anomaly through an MIL framework by learning a ranking model that predicts high anomaly scores for anomalous video segments.
	2019	IBL [41]	UC	Used an inner bag loss for MIL to increase the gap between the lowest and highest scores in a positive bag and reduce it in a negative one.
		GCN [45]	P, S, UC	A GCN was used to clean label noise, to directly apply fully supervised action classifiers to weakly supervised anomaly detection.
		Motion-Aware [44]	UC	Added temporal context to the MIL ranking model by using an attention block; the attention weights helped to identify anomalies better.
	2020	AR-Net [42]	S	A dynamic MIL loss enlarged the interclass dispersion; a centre loss reduced the intraclass distance of normal snippets.
	2021	MIST [49]	S, UC	Implemented a pseudo-label generator and an attention-boosted feature encoder to focus on anomalous regions.
		RTFM [46]	P, S, UC, X	A feature magnitude learning function was trained to recognise positive instances; self-attention mechanisms captured temporal dependencies.
		CRFD [50]	S, UC, X	Captured local-range temporal dependencies; enhanced features to the category space and further expanded the temporal modeling range.
		BD-LSTM [51]	UC, UC2L	A BD-LSTM network was used to reduce the inference time of the sequence of frames while maintaining competitive results.
	2022	MSL [47]	S, UC, X	Transformer-based MSL network to learn both video-level anomaly probability and snippet-level anomaly scores.
Fully-sup.	2019	Liu et al. [53]	UC	Implemented a region loss to explicitly drive the network to learn the anomalous region; a meta learning module prevented severe overfitting.
		Landi et al. [54]	UC2L	Considered spatiotemporal tubes instead of whole-frame video segments; existing videos were enriched with spatial and temporal annotations to allow bounding box supervision in both its train and test set.
	2022	UBnormal [55]	S, UC	Proposed the translation of simulated objects from its dataset to others using a CycleGAN, increasing performance.

3. Publicly Available Datasets

Generally, research on anomaly detection in video sequences has intensely focused on analysing video surveillance footage of pedestrians and crowds. Therefore, most of the available datasets concern those kinds of scenarios. Nonetheless, new datasets have tried to cover new areas, such as violent and criminal behaviours, and the surveillance of streets shared by pedestrians and vehicles. The current state of synthetic datasets will be covered as well as the possibility of expanding them to new use cases, for instance, to detect left behind objects and abnormal behaviour inside vehicles.

3.1. Real-World Datasets

3.1.1. Pedestrians and Crowds

The UMN dataset [58] was one of the first datasets used for benchmarking tasks of video anomaly detection. However, it had several limitations that led to a loss of relevance in recent benchmarks, mainly the existence of a single anomaly type. UMN is composed of 11 short clips of three indoor or outdoor scenes in which people suddenly start running away. There is no clear split between training and testing frames, and anomalies are only temporally labelled. Subway [57] has been declining in popularity for similar reasons. This indoor dataset is divided into two separate instances: Entrance and Exit. Only two long videos are provided, and there is uncertainty about which frames should be labelled as anomalous and used for the training and testing processes. The number of types of anomalies is reduced and include people jumping over the turnstiles or walking in the wrong direction.

The most widely used video anomaly detection dataset is the UCSD pedestrian dataset [12]. It was captured by a stationary video camera, focusing on two pedestrian walkways. This dataset contains two separate subsets: Ped1 and Ped2. The former is composed of 34 training videos and 36 testing videos, whilst the latter consists of 16 training video clips and 12 testing ones. The abnormal behaviours are connected to the presence of vehicles such as cars and bikers, as illustrated by Figure 3a,b. The CUHK Avenue dataset [11] is very common in benchmarks, and it was also acquired using a stationary video camera in the CUHK campus avenue. It has 16 training video samples and 21 test video samples. The abnormal behaviour represented in the scenes is connected to human actions, showing people littering items, walking on the grass, and throwing or abandoning objects in the background. However, both datasets possess severe limitations regarding their single-scene representation, lack of abnormality diversity, and amount of sequences. It is desirable to learn an anomaly detection model capable of performing well under multiple scenes and viewing angles. To address these drawbacks, ShanghaiTech [4] was developed, taking advantage of 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 camera angles, as Figure 3c,d exemplify. It contains 130 abnormal events and annotations for pixel-level ground truth of abnormal events.

Figure 3. Abnormal frames extracted from widely used datasets for training and benchmarking video anomaly tasks. (**a**) Two bikers amongst the pedestrians in Ped1 [12] dataset. (**b**) Car and biker in a pedestrian walkway in Ped2 [12] dataset. (**c**) A normal frame from ShanghaiTech dataset [4]. (**d**) Two people fighting in ShanghaiTech dataset [4] .

Recently, a new dataset, IITBCorridor [60], was proposed to tackle the lack of diversity of abnormal human activity. The videos were captured on the IIT Bombay campus, using a single-camera setup. There is a single scene, consisting of a corridor where walking and standing are considered regular activities. Protesting, fighting, and abandoning suspicious objects were some of the ten enacted activities. These activities extend from a single person to group-level anomalies, and frame-level labels were provided for training and validating models. Additionally, class labels for each abnormal activity were also provided for classification purposes. Another work that follows this path is ADOC [61], with data

acquired from a surveillance camera deployed on a large university campus. The dataset was created from video captured for 24 consecutive hours. The captured video encapsulates varying illumination conditions, with daytime and nighttime sequences, and crowded scenarios with background clutter. The data are annotated with 875 events.

3.1.2. Real-World Anomalies

Motivated by the limitations of previous datasets, UCF-Crime [40] was developed as a new large-scale dataset to evaluate video anomaly detection. It is composed of 1900 untrimmed videos of real-world surveillance footage, extracted from the internet, with an average length of 4 min each. It includes 13 types of anomalous events with a high impact on public safety, such as abuse, burglary, shoplifting and shooting. A comparison between a normal frame and a shooting sequence is displayed in Figure 4a,b. UCF-Crime contains annotated bounding boxes of anomalous regions in one image per 16 frames of each abnormal video. A considerable amount of available data was essential for the development of weakly supervised strategies.

XD-Violence [59] was originally released to develop a large-scale and multi-scene dataset for violence detection and classification. However, this task can be understood as a subset of anomaly detection, and the dataset can be used to benchmark new video anomaly detection methods. Furthermore, it contains audio-visual signals, allowing for the research on multi-modal solutions for this problem. XD-Violence consists of 4754 weak-labelled untrimmed videos with audio, which were collected from both films and YouTube. This dataset embraces a variety of scenarios and anomalies, for instance, rioting, car accidents and explosions, as shown in Figure 4c,d.

Figure 4. Comparison between normal and abnormal frames extracted from real-world anomalies datasets. (**a**) Frame from a normal activity extracted from UCF-Crime [40]. (**b**) Abnormal frame from UCF-Crime [40], showing a shooting. (**c**) Frame from a normal activity in XD-Violence [59]. (**d**) Abnormal frame from XD-Violence [59], representing an explosion.

3.1.3. Traffic

Most datasets that involve traffic consist of dashcam videos or surveillance videos to support the development of systems capable of anticipating traffic accidents. However,

the scope of this paper concerns anomaly detection of human-related behaviours. The Street Scene dataset [62] consists of 46 training video sequences and 35 testing video sequences taken from a static camera looking down on a scene of a two-lane street with bike lanes and pedestrian sidewalks. All of the footage is composed of daytime sequences, and it does not contain staged anomalies. The testing sequences have a total of 205 anomalous events consisting of 17 different anomaly types, such as jaywalking, cars outside their lane, and loitering. Although weather conditions are similar in every sequence, the dataset is challenging due to the variety of simultaneous activities occurring, moving background (e.g., trees moving with the wind) and changing shadows.

3.2. Synthetic Alternatives

UBnormal [55] is a novel supervised open-set benchmark composed of multiple virtual scenes for video anomaly detection. The artificial generation of the scenes of this dataset using Cinema4D is essential to provide pixel-level annotations for abnormal events in the training set, allowing for the use of fully supervised methods for video anomaly detection. The dataset consists of 29 virtual scenes with 660 anomalies and nine different types of abnormal behaviour. The videos were generated at 30 FPS and are composed of photorealistic frames as far as both background and actors are concerned. Normal and abnormal frames of some scenes are illustrated in Figure 5a,b, respectively. Additionally, the work of Acsintoae et al. [55] introduces an interesting concept of expanding real-world datasets. The translation of simulated objects from UBnormal to Avenue [11] or ShanghaiTech [4] was proposed using a CycleGAN [56], producing enhanced results when used to train state-of-the-art methods. Similar hybrid strategies could be studied as a solution to the lack of available public datasets for anomaly detection inside vehicles.

SVIRO-Uncertainty [63] is a high-quality synthetic dataset that is not directly related to the task of anomaly detection, which is understood as abnormal actions perpetrated by people. Nonetheless, it has the potential to be adapted to study a subset of this problem: the detection of abandoned or dangerous objects. SVIRO-Uncertainty is made up of sequences of the rear bench of a vehicle, in which each of the three seats might contain a passenger or an object, as displayed in Figure 5c,d. The original goal of this dataset was to train models capable of classifying the object that is occupying each position. However, it could be applied to abandoned object detection based on the context of its presence in the sequence. The dataset is quite large, containing two separate training sets, 4384 scenes with adult passengers only and 3515 using adults, child seats and infant seats. The adaptation of the dataset relies on providing context clues about the objects present in the frame, defining the anomaly as an object that should not be present in the sequence. An object could have been abandoned (e.g., a phone left in a seat) if its presence is not expected without a person present in the frame. Moreover, a relevant type of anomaly is linked to dangerous objects, such as firearms or knives. Due to the size of SVIRO-Uncertainty, providing such labels to a significant amount of sequences might be a virtually impossible task. A brief summary of the analysed datasets is available in Table 5.

Figure 5. Comparison between normal and abnormal frames extracted from synthetic datasets. (**a**) Frame from normal sequence in UBnormal [55], showing pedestrians walking down a street. (**b**) Frame from abnormal sequence in UBnormal [55]. Anomalous region emphasised with red contour. (**c**) Frame extracted from SVIRO-Uncertainty [63]. Expected behaviour of a passenger with their belongings. (**d**) Frame extracted from SVIRO-Uncertainty [63]. Objects unexpectedly abandoned in the back seat.

Table 5. Brief summary of relevant information regarding the analysed anomaly detection datasets. Legend: †—Data computed by Acsintoae et al. [55] based on the tracks from Georgescu et al. [64].

Dataset	Number of Frames			Scenes	Number of Anomalies		Resolution (px)
	Normal	Abnormal	Total		Types	Total	
UMN [58]	6165	1576	7741	3	1	11	320 × 240
Subway [57]	192,548 †	16,603 †	209,151	2	5	65 †	512 × 384
UCSD Ped1 [12]	9995	4005	14,000	1	5	54	238 × 158
UCSD Ped2 [12]	2924	1636	4560	1	5	23	320 × 240
CUHK Avenue [11]	26,832	3820	30,652	1	5	47	640 × 360
ShanghaiTech [4]	300,308	17,090	317,398	13	11	130	-
IITB-Corridor [60]	375,288	108,278	483,566	1	10	-	-
ADOC [61]	162,093	97,030	259,123	1	25	721	1920 × 1080
Street Scene [62]	159,341	43,916	203,257	1	17	205	1280 × 720
UCF-Crime [40]	-	-	13,741,393	1900	13	-	-
XD-Violence [59]	-	-	-	4754	6	-	-
UBnormal [55]	236,902	147,887	89,015	29	22	660	1080 × 720

4. Projects and Resources

4.1. Projects

Recently, several international projects intending to bridge the gap between research and the market for the next generation of security solutions have been launched. These projects do not focus solely on video surveillance of public spaces, one example being PREVENT PCP, which was launched to procure innovative and advanced systems to support security in public transport. This project involves some of the biggest transport operators in Europe, which consider that a concerted effort is required to develop solutions

in this area. The identification of unattended objects was chosen as the main challenge in providing safety in public transport, after the findings of PREVENT CSA, which served as a case study for this project. The detection of abandoned or suspect objects is a part of the broader anomaly detection problem, and the proposed scenario includes in-vehicle monitoring. The final goal of this project funded by the European Union's Horizon 2020 programme is to allow buyers to steer the development of cost-effective solutions directly toward their needs. Additionally, it reinforces the competitiveness of the EU technology and industrial base by funding projects of public interest.

European Union's Horizon 2020 programme is also responsible for the funding of i-DREAMS, which is a project designed to set up a framework for the definition, development, testing and validation of a context-aware *Safety Tolerance Zone* for driving. This analysis is meant to prevent drivers from getting too close to the boundaries of unsafe operation and to bring them back into the safety tolerance zone while driving. Although the full scope of the project transcends the recognition of the abnormal activity of the passengers, the detection of internal distractions, extreme emotions or health concerns might be supported by deep anomaly detection solutions.

A more traditional project, funded by the same programme, can be found in SecureIT. One of the explored domains consists of public space protection, especially at major events. Some important challenges are proposed, such as gathering and managing real-time information from multiple sources, since the targeted end-users are cities. The proposed solutions could later be transposed to other domains of interest.

4.2. Resources

Despite their compilation efforts, most surveys on deep anomaly detection for various scenarios only provide some very high-level outlines of the application conditions of the suggested models and a superficial review of the available datasets [1,65]. Although these are useful tools to deepen the theoretical knowledge of anomaly detection strategies, the limitations imposed by the accessible data must be explored. Moreover, this exploration effort provides an overview of the conditions required for improving and finding new application scenarios for the proposed methods. The work of Peng et al. [6] is as of yet the most detailed overview of the approaches taken by the current methods and their underlying intuitions. This is achieved by providing a comprehensive literature review while inspecting the current challenges and studying future opportunities for application. Furthermore, a collection of publicly accessible source codes and a large number of real-world datasets were presented by the authors. However, the work of Peng et al. [6] is lacking application-oriented reviews of opportunities and challenges posed by the exploration of novel scenarios. To fully explore and understand the constraints and potential issues of a specific real-world scenario, rather than a generic formulation of a problem, the attention that only a dedicated review can offer is required.

The availability of the code used to develop the models is essential not only to correctly replicate them but also to better understand some of the concepts that were presented in a paper. Table 6 lists some details on the code shared by the authors of some of the surveyed methods. The data used to create this table were originally compiled by Feng et al. [39]. Every repository was verified to confirm its accessibility and if it contained sufficient information to be properly implemented without major adjustments. Some additional information was also gathered, such as the Machine Learning (ML) framework that was used. Regarding this last topic, there is an interesting tendency, evident by analysing Table 6, on the growing popularity of PyTorch as the preferred framework employed by researchers investigating video anomaly detection. This compilation is a starting point for selecting models that could be adapted to anomaly detection in new scenarios, such as monitoring passengers and objects inside vehicles.

Table 6. Available source code of the surveyed methods, based on the compilation made by Feng et al. [39]. Abbreviation: a.o.—accessed on date (format dd/mm/yy).

Type	Year	Method	Availability	ML Framework
Semi-supervised	2017	S-RNN [4]	Link, a.o. 01/08/22	TensorFlow
		ConvLSTM-AE [20]	Link, a.o. 01/08/22	Caffe
	2018	FFP [3]	Link, a.o. 29/07/22	TensorFlow
		ALOCC [66]	Link, a.o. 01/08/22	TensorFlow
	2019	Mem-AE [23]	Link, a.o. 01/08/22	PyTorch
		AMC [67]	Link, a.o. 01/08/22	TensorFlow
	2020	MNAD [36]	Link, a.o. 27/07/22	PyTorch
		OGNet [68]	Link, a.o. 01/08/22	PyTorch
		VEC [27]	Link, a.o. 01/08/22	PyTorch
Weakly supervised	2018	Sultani et al. [40]	Link, a.o. 01/08/22	Theano
	2019	GCN [45]	Link, a.o. 01/08/22	PyTorch
		MLEP [69]	Link, a.o. 01/08/22	TensorFlow
	2020	AR-Net [42]	Link, a.o. 01/08/22	PyTorch
		XD-Violence [59]	Link, a.o. 30/07/22	PyTorch
	2021	MIST [49]	Link, a.o. 29/07/22	PyTorch
		RTFM [46]	Link, a.o. 01/08/22	PyTorch

5. Challenges, Approaches and Opportunities for In-Vehicle Monitoring

Anomaly detection in confined spaces, such as the interior of vehicles, is an interesting new application scenario for these methods. However, as the work of Augusto et al. [5] demonstrates, the development of solutions for this use case is still fully dependent on the availability of private datasets. A subset of a dataset provided by Bosch Car Multimedia containing videos of nine different actor pairs performing various activities in the backseat of a vehicle was used by the authors. Every video featured two actors in every frame, and the anomalies present in the subset are strictly related to violent interactions between two individuals only (e.g., slapping and punching). However, the relevance of objects was not considered in this work, whether for representing a danger to the passengers or simply as an object that was left behind by one of them. The latter is of significant importance in the sugegsted shared autonomous vehicle scenario.

Creating new datasets or expanding existing ones appears to be an immediate need for considering new use applications for anomaly detection. The former is a complex and costly task that implies allocating resources for staging and recording the desired interactions. An additional bureaucratic effort is also required to obtain permission from the actors involved. Moreover, a post-recording labelling effort is time consuming. Hence, an attractive option relies on synthetic data that could be generated for direct use or to augment available data. The work of Acsintoae et al. [55] is referred to as an interesting approach to the translation of simulated objects to real-world datasets using a CycleGAN [56]. Similar hybrid strategies could be employed to circumvent the lack of data for in-vehicle monitoring applications. Furthermore, such strategies could pre-emptively add some artificial variety to the available video sequences. As it

was referred, the videos provided by Bosch Car Multimedia that were used by Augusto et al. [5] contained only nine different actor pairs. The work of Capozzi et al. [70] has linked the lack of actor independence with the underperformance of the trained models, as a bias is developed linking certain actors to certain actions, instead of learning the pattern of the action. Moreover, a larger pool of actors' characteristics, artificially increased or not, is essential to expose the model to diverse scenarios. Some additional challenges arise from the type and model of the vehicle that was used. For instance, the shape of the windows affects the background and light conditions in the captured scenes. Furthermore, the seats of the vehicle influence the range of movements of the passengers as well as their pose. The diversity of vehicles and actors are essential factors to produce a robust model.

Choosing the best model for a new use case such as anomaly detection inside of a vehicle is not straightforward. The typical scenario of the reviewed publicly available datasets does not faithfully represent the new environment in which anomalies must be detected; therefore, their use does not produce an authentic benchmark of the proposed methods. Most of these sequences were captured with stationary video cameras that were recording static backgrounds. Although cameras inside vehicles are also stationary, windows on a moving vehicle produce a partially moving background on the recorded sequence. The distance between the cameras and the subjects is much smaller inside a vehicle, increasing the effect of geometric distortions on the captured information. Additionally, headlights of other vehicles, public illumination and occlusions of sunlight produce more frequent illumination perturbations in the scene than those found on datasets that focus on a pedestrian walkway, for instance. The behaviour of the available models in such scenes is uncertain, as these did not have to specifically build and test tools for such problems. However, they cannot be neglected to build a successful application for this use case. Furthermore, in the available datasets, especially the ones regarding pedestrians and crowds, the entirety of the body of the actors is visible; therefore, the models can benefit from this information to detect anomalies. However, inside confined spaces, this might not be possible. Taking into consideration the in-vehicle scenario, due to the limited available camera positions, part of the legs of the passengers are occluded, as Figure 5c demonstrates. Therefore, in such tasks, the models are limited to partial information regarding the human actors and the area in which the actions take place.

None of the datasets that were analysed in Section 3 present a convenient tool for training and benchmarking a model for anomaly detection inside a vehicle or similarly confined spaces. The datasets that comprise sequences of pedestrians and crowds were mostly recorded outdoors and cover a great area when compared to the new scenario of interest. Additionally, the normal samples that these sequences present consist of people walking or simply standing, which are actions that would be considered abnormal inside a car. As far as confined spaces are concerned, the datasets that present real-world anomalies, UCF-Crime [40] and XD-Violence [59], possess some scenes that fit this context. However, the available labels do not give any information regarding the location in which the sequences take place; therefore, an additional labelling effort would be required. Moreover, they do not present coherence in terms of the placement of the camera or the type of confined spaces presented, as these videos were extracted from films or the internet. On the other hand, SVIRO-Uncertainty [63] depicts an in-vehicle scenario, despite not presenting relevant information for anomaly detection in terms of abnormal actions perpetrated by the passengers. Its potential remains solely linked to the detection of dangerous or abandoned objects, which is a subset of anomaly detection.

A common issue with the proposed deep anomaly detection techniques was noted by Pang et al. [6]. Most anomaly detection studies focus on detection performance only, ignoring the capability of illustrating the identified anomalies. Although it would be relevant to classify the abnormal behaviour that was detected, the detection could represent a novel anomaly. Hence, it is crucial to at least provide spatial cues that demonstrate the specific data portion that is anomalous. These cues might prove useful as a tool for interpreting such complex models and identifying scenarios in which they could be missing. Furthermore, the works of Liu et al. [53] and Landi et al. [54] have proven that locality

is a powerful instrument to improve performance and reduce background bias. Some re-labelling was required to construct both attention-driven models, but robust results were achieved. Additionally, the model proposed by Landi et al. [54] was able to provide spatiotemporal proposals for unseen surveillance videos leveraging only video-level labels, which is a useful feature for the needed expansion of datasets.

6. Conclusions

In this article, various deep learning methods for anomaly detection in videos were discussed. Studying the defining characteristics of state-of-the-art methods is important not only to gain a better understanding of the general problem of anomaly detection but also to understand how the offered solutions could fit into the new scenario of interest: in-vehicle monitoring. The major contributions of the analysed works are briefly summarised in Table 4. Additionally, Table 6 provides a compilation of the available source code; the code present in these repositories comprises an interesting starting point for replicating and improving these models for new applications.

The analysis of state-of-the-art techniques provided a deeper understanding of the background of these models and its influence on their current limitations regarding in-vehicle monitoring. The focus on crowded scenes and outdoor spaces led to a failure to consider problems associated with the nature of this new scenario. For instance, the surveillance of Shared Autonomous Vehicles must consist of a much closer recording of the subjects, which raises questions about the importance of actor independence and the effect of geometric distortions on the captured information caused by the lens of the camera. Moreover, these models assume a mostly static background, although the movement of the car and the presence of windows result in moving backgrounds. Additionally, frequent illumination changes (e.g., a cloud covering the sun) result in a more intense impact on the visual information in such scenarios.

The main limitation of the implementation of anomaly detection solutions to in-vehicle monitoring is the lack of data samples explicitly dedicated to the detection of abnormal behaviours inside a vehicle or similarly confined spaces. Hence, there are currently no public datasets that could be directly used as a tool for training and benchmarking such models. The development of solutions for this use case is still fully dependent on the availability of private datasets. Although newer datasets have been adapted to benchmark models created for anomaly detection tasks, their original focus was related to action recognition tasks. The reviewed synthetic datasets presented high-quality images with an interesting amount of annotations but do not comprise a compatible set of data instances for this task. Although this is a severe challenge, it also provides a great opportunity to study techniques for data augmentation and generation. In this paper, several approaches were proposed to be implemented in future works. They can be summarised in a two-stage process. Firstly, available datasets, mainly the ones covering diverse real-world anomalies, must be extensively studied to find instances representative of the real-world settings that the systems will face, providing an initial reference. Secondly, the expansion and adaptation of similar instances should be contemplated. This could be achieved through the translation of simulated objects or actors to tackle the lack of available sequences but also their reduced diversity of actors, actions and significant illumination changes. This approach could also reduce the labelling effort of new captures.

This review initiates an important discussion on application-oriented issues related to deep anomaly detection for in-vehicle monitoring, which is a field that presents a high potential for exploration in future works. Other surveys and reviews have disregarded this scenario and its specificities, despite its relevance, as shown by the funded projects, such as Prevent PCP, that aim to take advantage of innovative solutions for applying anomaly detection to this scenario. Moreover, in-vehicle monitoring increases the interest in the optimisation effort of anomaly detection models for embedded systems, as its implementation requires the capability of running locally in resource-limited hardware.

Author Contributions: Conceptualisation by F.C., P.C. and J.C.; funding acquisition, supervision, and writing—review and editing by P.C. and J.C.; investigation, data collection, formal analysis, and writing—original draft preparation by F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Funds through the Portuguese funding agency, FCT—Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data sources are publicly available and mentioned in the references.

Acknowledgments: The authors acknowledge the support provided by Bosch Car Multimedia Portugal in Project Aurora.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Kiran, B.R.; Thomas, D.M.; Parakkal, R. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *J. Imaging* **2018**, *4*, 36. [CrossRef]
2. Xu, D.; Yan, Y.; Ricci, E.; Sebe, N. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* **2017**, *156*, 117–127. [CrossRef]
3. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection–a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.
4. Luo, W.; Liu, W.; Gao, S. A revisit of sparse coding based anomaly detection in stacked rnn framework. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 341–349.
5. Augusto, P.; Cardoso, J.S.; Fonseca, J. Automotive interior sensing-towards a synergetic approach between anomaly detection and action recognition strategies. In Proceedings of the 2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS), Virtual Event, 9–11 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 162–167.
6. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep learning for anomaly detection: A review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, pp. 1–38. Article 38 [CrossRef]
7. Zimek, A.; Schubert, E.; Kriegel, H.P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat. Anal. Data Min. ASA Data Sci. J.* **2012**, *5*, 363–387. [CrossRef]
8. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI—Explainable artificial intelligence. *Sci. Robot.* **2019**, *4*, eaay7120. [CrossRef]
9. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, pp. 1–58. Article 15 [CrossRef]
10. Pang, G.; Shen, C.; Jin, H.; Hengel, A. Deep weakly-supervised anomaly detection. *arXiv* **2019**, arXiv:1910.13601.
11. Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2720–2727.
12. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1975–1981.
13. Wang, X.; Che, Z.; Jiang, B.; Xiao, N.; Yang, K.; Tang, J.; Ye, J.; Wang, J.; Qi, Q. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, pp. 2301–2312 [CrossRef]
14. Xu, D.; Ricci, E.; Yan, Y.; Song, J.; Sebe, N. Learning deep representations of appearance and motion for anomalous event detection. *arXiv* **2015**, arXiv:1510.01553.
15. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
16. Sabokrou, M.; Fathy, M.; Hoseini, M. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electron. Lett.* **2016**, *52*, 1122–1124. [CrossRef]
17. Zhao, Y.; Deng, B.; Shen, C.; Liu, Y.; Lu, H.; Hua, X.S. Spatio-temporal autoencoder for video anomaly detection. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1933–1941.
18. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
19. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
20. Luo, W.; Liu, W.; Gao, S. Remembering history with convolutional lstm for anomaly detection. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 439–444.

21. Ranzato, M.; Szlam, A.; Bruna, J.; Mathieu, M.; Collobert, R.; Chopra, S. Video (language) modeling: A baseline for generative models of natural videos. *arXiv* **2014**, arXiv:1412.6604.
22. Wisdom, S.; Powers, T.; Pitton, J.; Atlas, L. Interpretable recurrent neural networks using sequential sparse recovery. *arXiv* **2016**, arXiv:1611.07252.
23. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; 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, Seoul, Korea, 27 October–2 November 2019; pp. 1705–1714.
24. Ravanbakhsh, M.; Nabi, M.; Sangineto, E.; Marcenaro, L.; Regazzoni, C.; Sebe, N. Abnormal event detection in videos using generative adversarial nets. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1577–1581.
25. Ganokratanaa, T.; Aramvith, S.; Sebe, N. Unsupervised anomaly detection and localization based on deep spatiotemporal translation network. *IEEE Access* **2020**, *8*, 50312–50329. [CrossRef]
26. 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, Nice, France, 21–25 October 2019; pp. 1805–1813.
27. Yu, G.; Wang, S.; Cai, Z.; Zhu, E.; Xu, C.; Yin, J.; Kloft, M. Cloze test helps: Effective video anomaly detection via learning to complete video events. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 583–591.
28. Chen, C.; Xie, Y.; Lin, S.; Yao, A.; Jiang, G.; Zhang, W.; Qu, Y.; Qiao, R.; Ren, B.; Ma, L. Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection. In Proceedings of the American Association for Artificial Intelligence,Virtual, 22 February–1 March 2022; pp. 1–9.
29. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
30. Georgescu, M.I.; Barbalau, A.; Ionescu, R.T.; Khan, F.S.; Popescu, M.; Shah, M. Anomaly detection in video via self-supervised and multi-task learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12742–12752.
31. Lee, S.; Kim, H.G.; Ro, Y.M. BMAN: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Trans. Image Process.* **2019**, *29*, 2395–2408. [CrossRef] [PubMed]
32. Bergmann, P.; Fauser, M.; Sattlegger, D.; Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4183–4192.
33. Ionescu, R.T.; Khan, F.S.; Georgescu, M.I.; Shao, L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7842–7851.
34. Doshi, K.; Yilmaz, Y. Any-shot sequential anomaly detection in surveillance videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 934–935.
35. Doshi, K.; Yilmaz, Y. Continual learning for anomaly detection in surveillance videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020, pp. 254–255.
36. Park, H.; Noh, J.; Ham, B. Learning Memory-guided Normality for Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14372–14381.
37. Cai, R.; Zhang, H.; Liu, W.; Gao, S.; Hao, Z. Appearance-motion memory consistency network for video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence,Virtually, 2–9 February 2021; Volume 35, pp. 938–946.
38. Ramachandra, B.; Jones, M.; Vatsavai, R. Learning a distance function with a Siamese network to localize anomalies in videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2598–2607.
39. Feng, J.C. Papers for Video Anomaly Detection, Released Codes Collection, Performance Comparision. 2022. Available online: https://github.com/fjchange/awesome-video-anomaly-detection (accessed on 26 July 2022).
40. Sultani, W.; Chen, C.; Shah, M. Real-world Anomaly Detection in Surveillance Videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488.
41. Zhang, J.; Qing, L.; Miao, J. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 4030–4034.
42. Wan, B.; Fang, Y.; Xia, X.; Mei, J. Weakly supervised video anomaly detection via center-guided discriminative learning. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), Virtual, 6–10 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
43. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
44. Zhu, Y.; Newsam, S. Motion-aware feature for improved video anomaly detection. *arXiv* **2019**, arXiv:1907.10211.

45. Zhong, J.X.; Li, N.; Kong, W.; Liu, S.; Li, T.H.; Li, G. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,Long Beach, CA, USA, 15–20 June 2019; pp. 1237–1246.
46. 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 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 4975–4986.
47. Li, S.; Liu, F.; Jiao, L. Self-training multi-sequence learning with Transformer for weakly supervised video anomaly detection. In Proceedings of the AAAI, Virtual, 22 February–1 March 2022 ; Volume 24.
48. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 3202–3211.
49. Feng, J.C.; Hong, F.T.; Zheng, W.S. MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
50. Wu, P.; Liu, J. Learning causal temporal relation and feature discrimination for anomaly detection. *IEEE Trans. Image Process.* **2021**, *30*, 3513–3527. [CrossRef]
51. 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]
52. Ullah, W.; Ullah, A.; Hussain, T.; Muhammad, K.; Heidari, A.A.; Del Ser, J.; Baik, S.W.; De Albuquerque, V.H.C. 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]
53. Liu, K.; Ma, H. Exploring background-bias for anomaly detection in surveillance videos. In Proceedings of the 27th ACM International Conference on Multimedia,Nice, France, 21–25 October 2019; pp. 1490–1499.
54. Landi, F.; Snoek, C.G.; Cucchiara, R. Anomaly locality in video surveillance. *arXiv* **2019**, arXiv:1901.10364.
55. Acsintoae, A.; Florescu, A.; Georgescu, M.; Mare, T.; Sumedrea, P.; Ionescu, R.T.; Khan, F.S.; Shah, M. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022.
56. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
57. Adam, A.; Rivlin, E.; Shimshoni, I.; Reinitz, D. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 555–560. [CrossRef] [PubMed]
58. Unusual Crowd Activity Dataset of University of Minnesota. 2006. Available online: http://mha.cs.umn.edu/movies/crowdactivity-all.avi (accessed on 2 August 2022).
59. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; 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*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 322–339.
60. Rodrigues, R.; Bhargava, N.; Velmurugan, R.; Chaudhuri, S. Multi-timescale trajectory prediction for abnormal human activity detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Village, CO, USA, 1–5 March 2020; pp. 2626–2634.
61. Pranav, M.; Zhenggang, L.; et al. A day on campus—An anomaly detection dataset for events in a single camera. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–December 2020.
62. Ramachandra, B.; Jones, M. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 2569–2578.
63. Dias Da Cruz, S.; Taetz, B.; Stifter, T.; Stricker, D. Autoencoder Attractors for Uncertainty Estimation. In Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022.
64. Georgescu, M.I.; Ionescu, R.; Khan, F.S.; Popescu, M.; Shah, M. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4505–4523. [CrossRef] [PubMed]
65. Chalapathy, R.; Chawla, S. Deep learning for anomaly detection: A survey. *arXiv* **2019**, arXiv:1901.03407.
66. Sabokrou, M.; Khalooei, M.; Fathy, M.; Adeli, E. Adversarially learned one-class classifier for novelty detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3379–3388.
67. Nguyen, T.N.; Meunier, J. Anomaly detection in video sequence with appearance-motion correspondence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1273–1283.
68. Zaheer, M.Z.; Lee, J.h.; Astrid, M.; Lee, S.I. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14183–14193.
69. Liu, W.; Luo, W.; Li, Z.; Zhao, P.; Gao, S. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, 10–16 August 2019; pp. 3023–3030.
70. Capozzi, L.; Barbosa, V.; Pinto, C.; Pinto, J.R.; Pereira, A.; Carvalho, P.M.; Cardoso, J.S. Towards Vehicle Occupant-Invariant Models for Activity Characterisation. *IEEE Access* 2022, *accepted*. [CrossRef]