electronics

MDPI

Review

Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance

Sabah Abdulazeez Jebur [1,2], Khalid A. Hussein [3], Haider Kadhim Hoomod [3], Laith Alzubaidi [4,*] and José Santamaría [5]

[1] Department of Computer Sciences, University of Technology, Baghdad 00964, Iraq
[2] Department of Computer Techniques Engineering, Imam Al-Kadhum College (IKC), Baghdad 00964, Iraq
[3] Department of Computer Science, College of Education, Mustansiriyah University, Baghdad 00964, Iraq
[4] School of Mechanical, Medical, and Process Engineering, Queensland University of Technology, Brisbane, QLD 4000, Australia
[5] Department of Computer Science, University of Jaén, 23071 Jaén, Spain
* Correspondence: l.alzubaidi@qut.edu.au

Abstract: In the last few years, due to the continuous advancement of technology, human behavior detection and recognition have become important scientific research in the field of computer vision (CV). However, one of the most challenging problems in CV is anomaly detection (AD) because of the complex environment and the difficulty in extracting a particular feature that correlates with a particular event. As the number of cameras monitoring a given area increases, it will become vital to have systems capable of learning from the vast amounts of available data to identify any potential suspicious behavior. Then, the introduction of deep learning (DL) has brought new development directions for AD. In particular, DL models such as convolution neural networks (CNNs) and recurrent neural networks (RNNs) have achieved excellent performance dealing with AD tasks, as well as other challenging domains like image classification, object detection, and speech processing. In this review, we aim to present a comprehensive overview of those research methods using DL to address the AD problem. Firstly, different classifications of anomalies are introduced, and then the DL methods and architectures used for video AD are discussed and analyzed, respectively. The revised contributions have been categorized by the network type, architecture model, datasets, and performance metrics that are used to evaluate these methodologies. Moreover, several applications of video AD have been discussed. Finally, we outlined the challenges and future directions for further research in the field.

Keywords: deep learning; anomaly detection; human behavior; video surveillance

Citation: Jebur, S.A.; Hussein, K.A.; Hoomod, H.K.; Alzubaidi, L.; Santamaría, J. Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. *Electronics* **2023**, *12*, 29. https://doi.org/10.3390/electronics12010029

Academic Editor: Stefanos Kollias

Received: 1 December 2022
Revised: 17 December 2022
Accepted: 19 December 2022
Published: 22 December 2022

1. Introduction

The actions that can be observed in a person as a result of being exposed to an internal or external stimulus constitute what is known as human behavior. The environment, including social interactions, provides external stimuli, while internal stimuli include something like a person's ideas, memories, perceptions, or attitudes [1]. Abnormal behavior can therefore be defined as actions that are not expected to appear in a specific context. From a computer vision (CV) point of view, abnormality refers to data patterns that are skewed from normal data. It is also known as an anomaly, outlier, or novelty. It also may be termed unusual, irregular, atypical, inconsistent, unexpected, rare, erroneous, faulty, fraudulent, malicious, unnatural, or a strange activity [2,3]. Anomaly detection (AD), aka. target detection, for video streams is a vital domain in many important areas of CV, e.g., video surveillance, autonomous vehicles, robotics, virtual reality, smart cities, and medical imaging [4]. AD has been used in many areas of video surveillance to detect abnormal activities such as running, climbing, falling, fighting, and robbing [5–10]; violence, loitering, and vandalism [11–15]; personal intrusion [16], autism and drug addiction [17],

and reckless driving [18]. It is also used to detect abnormal behavior in specific places, such as petrol stations [19] and elevators [20]. As previously stated, an anomaly is an irregular scene in a particular time and place. For example, a crowd at a market on an average day is regarded as normal, while the same crowd in the same place during a curfew is considered abnormal. Similarly, a crowd at a marathon would not be regarded as an anomaly, but a crowd in front of a building would be. In other words, the definition of anomaly can evolve over time, and the current concept of normal or anomaly behavior might not be properly represented in the future. In addition, challenges such as diversity in scenarios, noisy videos, low probability of occurrence of anomalies, and infrequent and low availability of labelled data for anomalous activity, all of which makes AD a challenging task for Artificial Intelligence (AI) [21]. Usually, machine learning (ML) algorithms need reliable features to function properly in order to both characterize the input data and classify the output results. Therefore, correctly recognizing behaviors relies on well-designed features, which have a direct impact on classification accuracy. Classification accuracy may decrease if feature extraction is based on empirical experience. Unlike ML, deep learning (DL) uses neural network (NN) models to automatically identify and extract features from input data without requiring feature extraction stages. With the help of DL, a specific method is available for classifying data that can scale to enormous amounts of data and incorporate complex features. One of the main benefits of DL is that it eliminates the need for any kind of preprocessing before acquiring feature descriptions. During training, the NN may automatically determine a large number of unknown parameters. Training takes a lot of time, however the results achieved pay off in the end [22]. Finally, the success of DL methods tackling with the AD problem is due to the ability to extract valuable and complex features from videos using nonlinear transformations. Furthermore, these kinds of methods can detect anomalies in time and space. Here, the localization method finds each frame that is abnormal and explains which portion of this frame is unusual, while detection focuses on the video fragments that have anomalies across all videos [1,2]. DL models such as convolution neural networks (CNNs), auto-encoders (AEs), generative neural networks (GANs), and recurrent neural networks (RNNs) have achieved remarkable performance addressing the AD problem. This paper provides the following main contribution:

- Review of the most relevant state-of-the-art contributions in the last four years dealing with DL applied to the AD problem.
- Detailed categorization of the existing methods in AD by classifying the approaches according to the specific DL methods and the adopted architectural models for AD.
- A comprehensive analysis of the DL architectures used in AD has been introduced to make it easy for a researcher to choose which approach may be more appropriate for the particular AD application.
- A performance evaluation of methodologies are discussed in terms of datasets and measures of performance.
- A discussion of the current challenges and needs in the domain of DL applicable to AD is put forth.
- A description of those new trends in DL-based AD are discussed to provide several interesting ideas to be considered in future research.

This review is organized as follows: Section 2 introduces the classification of anomalies in video streaming. Section 3 deals with DL methods for AD. Section 4 presents various architectural models that are utilized for AD. Benchmarked datasets and performance metrics are reviewed in Sections 5 and 6, respectively. Several applications and research challenges in DL-based AD approaches are discussed in Sections 7 and 8, respectively. In addition, this paper outlined the future direction and research opportunities in Section 9. The conclusions of our work are drawn in Section 10.

Survey Methodology

The majority of significant research papers that have been reviewed in this article were published during 2019–2022. The main focus was papers from the most reputed

publishers, such as IEEE, Springer, ArXiv, Elsevier, and MDPI. We have reviewed more than 100 papers on the topic of anomaly detection in video. The most keywords used for search criteria for this review paper are ("Video Anomaly Detection" OR "Video Anomaly Recognition"), ("Abnormal Human Behavior"), ("Deep Learning" AND "Anomaly Detection"), ("CNN" and "Anomaly Detection"), ("Applications of Video Anomaly Detection"), and ("Challenges in Anomaly Detection").

2. Classifications of Anomaly Detection

Anomaly detection in video depends on the type of the anomalies. Hence, video anomalies may be classified as follows:

2.1. Image-Based and Video-Based Detection

Models used to detect anomalous behavior fall into two categories, depending on whether they rely on a single frame (image-based detection) or a sequence of frames (video-based detection) [7,23,24]. The first model processes still image frames to extract spatial (shape)features and detect the target's behavior, like falling or using a weapon. However, this model frequently leads to misjudgments because it cannot effectively extract the temporal (time) information of the behavior. Without temporal information, it is difficult to recognize the motion of an action. For example, it is hard to identify the difference between touching someone's cheek and slapping them in a single frame. The second model processes several consecutive frames to extract both spatial and temporal information in the video to classify target behaviors like violence, robbery, and fighting. This model has scored high accuracy as well as flexibility. Therefore, it has been widely studied and applied in video-based human behavior analysis systems.

2.2. Single-Point Anomaly and Group Anomalies

Based on how many anomalies there are in a scene, anomaly events can be divided into two categories: single point anomalies and group anomalies. In [25], a single point anomaly is described as an anomalous activity of an individual entity, also known as an entity-based anomaly. In other words, the data points that significantly differ from the rest of the data points are what are known as point-anomalies, and some examples of such behaviors include loitering. Interaction anomaly [26], also known as group-based anomaly, is the unusual interaction of groups of entities. Examples include fighting people or car accidents. From single entity-based anomalies to interaction-based anomalies, the complexity and time requirements for anomaly detection and localization increase.

3. DL Methods of AD

The nature of the input data is the basic factor that determines which deep neural network (DNN) is employed in the AD task. DL techniques used for AD can be classified into the following categories based on the extent of availability of labels: (1) Supervised video AD, (2) Semi-supervised video AD, (3) Unsupervised video AD, (4) Transfer learning-based video AD. (5) Deep active learning-based video AD. (6) Deep reinforcement Learning-based AD, and (7) Deep hybrid models.

3.1. Supervised Learning-Based AD for Video Streaming

These techniques utilize annotated data as input to train the model. Supervised DL networks start with initial parameters, then these parameters are repeatedly updated using a back-propagation algorithm to get an improved estimate for the preferred results [27]. Broadly, supervised DL-based classification models consist of feature extraction networks and classification networks. Specifically, supervised networks in AD applications involve the training of two types of classifiers: a binary classifier and a multiclass classifier. The binary classifier uses labeled data of normal and abnormal samples, whereas, with a multi-class classifier, the training data contains labeled instances of the normal class and multiple abnormal classes. There are multiple supervised learning techniques that are used in AD,

such as CNNs and RNNs. In addition, the CNN category includes VGG16, VGG19, and Yolo series algorithms. While the RNN category includes long-short-term memory (LSTM) and gated recurrent units (GRUs) algorithms [27,28]. The key advantage of these techniques is the ability to gather data and produce data output from prior knowledge. Furthermore, they are simpler and have higher performance compared to other DL techniques. On the other hand, the disadvantage of these techniques is that decision boundaries might be overstrained when the training set lacks examples that belong to a class. In addition, they require precise labels for a variety of normal and anomalous classes, which are often not available [25].

3.2. Semi-Supervised Learning-Based AD for Video Streaming

In the semi-supervised technique, the training process utilizes datasets with weakly labeled instances where it is assumed that all training instances have a single class label. Here, this technique develops a discriminative boundary around the normal instances. Thus, test instances are marked as anomalous when they do not belong to the majority class. This method takes advantage of both supervised and unsupervised methods. Generative adversarial networks (GANs), GRUs, and LSTMs are used for semi-supervised learning in AD models. One of the advantages of this technique is that it requires the least amount of labeled data. On the other hand, the main drawback of this technique is that it is susceptible to the over-fitting issue and that irrelevant input features found in training data may result in wrong classification [25,27].

3.3. Unsupervised Learning-Based AD for Video Streaming

This technique allows one to perform the learning process in the absence of labeled data. It learns the inherent data features such as distance or density to distinguish between normal and abnormal data to facilitate AD by finding commonalities within the data. DL techniques such as Auto Encoders (AEs), Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBNs), and GANs, as well as GRUs and LSTM algorithms, have been used for developing unsupervised learning-based models in AD applications. The success of the unsupervised video AD methods depends on the availability of large video datasets and high computational resources. One of the benefits of this method is that it is a low-cost way to find outliers because the algorithms do not need to be trained with annotated data [25,27].

3.4. Transfer Learning-Based AD for Video Streaming

In general, there are two ways to train NNs: scratch learning (SL) and transfer learning (TL). In SL, a network starts to learn with random initial weights. This type of learning needs a large amount of data, powerful computing power, and a very long processing time. To address these challenges, the concept of TL has been proposed as a means of overcoming these difficulties. TL is a technique that allows you to use the knowledge gained from one task to improve the performance of a related task. This is done by using the weights and parameters learned from the first task as a starting point for the second task. This can help reduce the amount of training data and computational resources required to achieve good performance [4,29–32].

3.5. Deep Active Learning-Based AD for Video Streaming

Active learning is the process by which an algorithm repeatedly asks human annotators for labels to enhance training and improve performance. Broadly, active learning analyzes data sets and makes the assumption that values for updating the model vary among samples within the same data set. High-scoring examples are prioritized for inclusion in the training set. This is done to improve the performance of the model, reduce the number of false positives, and lower the cost of labeling. Active learning reduces the ambiguous nature of anomalies in the AD framework by introducing suitable priors with the assistance of a domain expert. Moreover, by simply demanding a minimal number of labels to boost

model performance, it also tackles the issues of an unbalanced dataset, concept drift—where data patterns are constantly changing—and the high labeling cost [33,34].

3.6. Deep Reinforcement Learning-Based AD

A strategy based on DRL actively looks out for novel classes of anomalies that fall outside the purview of the labeled training data. This strategy learns to strike a balance between taking advantage of its current data model and looking for new classes of anomalies. As a result, it can use the labeled anomaly data to increase the detection accuracy without restricting the set of anomalies it searches for to the examples provided. The general concept behind utilizing RL for decision-making issues is that an agent will be able to learn from the environment by interacting with it and obtaining rewards for certain actions; this concept derives from humans' natural method of learning via their experiences. DRL combines RL and DL to solve more complicated problems. Solving such challenges involves dealing with high-dimensional data and environments, sparse reward signals, and uncertainties in the agent's observations. In particular, the DRL-based AD approach leverages the labeled anomaly data to increase the detection accuracy without restricting the set of anomalies sought to those provided as anomalous examples. The approach accomplishes this by interacting with an environment created using the training data [35,36].

3.7. Deep Hybrid Models-Based AD for Video

The hybrid models combine multiple models together in such a way as to improve AD for video streaming. These models also work well with input data that has high dimensions, such as video data. Deep hybrid models mostly use DNNs as a feature extraction process and traditional ML algorithms to detect anomalous activities. For instance, ref. [4] put forth a hybrid approach for AD in video based on CNN and SVM. CNN is used to extract descriptive features. Next, the feature vector is passed into a binary SVM to construct the abnormal event detection model. While ref. [37] used 3D-ConVNet and AEs methods, 3D-ConVNet to learn video representation automatically and extract features from both spatial and temporal dimensions, and AEs to predict the future frames. Next, the anomaly score is calculated based on the reconstruction error. Another example, ref. [12] introduced a model integrating two methods, GAN and multi-instance learning (MIL), in a single framework to predict future anomalies. GAN for future frame prediction and MIL for anomalies detection.

4. DL Architectures for AD

A DNN's architecture specifies its layering, width, depth, and node types. Many network structures have been proposed for extracting features and identifying actions. For behavior recognition in videos, DL networks need to take into account more than just spatial information extraction, as is the case with image-based systems. Without temporal information, the motion of an action cannot be differentiated; for example, the act of opening a door is similar to that of closing a door. Action recognition in video can be improved by making use of temporal motion data. Clearly, there is a connection between temporal motion data and video action detection. Table 1 presents state-of-the-art DL methods used in anomaly detection in videos during the last three years. The DL architectures used to find video anomalies can be roughly put into the following groups:

Table 1. DL approaches applied to AD for video.

[Ref.], Year	Type of Network	Proposed Architecture	Dataset (Accuracy)	Examples of Anomalies
[5], 2020	CNN	human skeleton, YOLOv3, Multi-scale information fusion network	UVF-101, HMDN51, and camera (96.3%)	Run, fall, fight

Table 1. *Cont.*

[Ref.], Year	Type of Network	Proposed Architecture	Dataset (Accuracy)	Examples of Anomalies
[4], 2020	CNN	VGGNet-19 pretrained network, binary SVM	UMN (97.44%), UCSD-ped1 (86.69%)	Carts, bikers, skateboarders, running, person walking over the grass
[17], 2020	CNN, RNN	Combined CNN-RNN	NAHFE (89.5%)	Drug addiction, autism, criminal mentality.
[6], 2020	CNN	Canny edge detection algorithm, 3D-ConVNet	HMDB51 and Hollywood2 (93%)	Climbing, fighting, falling
[11], 2020	CNN, RNN	ConvLSTMs	Hockey (99%), Violent Flow (93.75%), RLV (96.74)	Violence
[18], 2020	CNN	YOLOv2	Camera (99.8%	Reckless driving
[21], 2021	LSTM, AEs	Convolution AE and sequence to sequence LSTM	UMN (87%)	Sudden running
[12], 2021	CNN, GAN	3D-ConVNet	CUHK Avenue (68.94%), ShanghaiTech (88.26%)	Crime
[8], 2021	CNN	3D-ConVNet	Behave (91.75%), Caviar (92.86%)	Robbery, fight
[10], 2021	GRU, FFN	Human skeleton, GRU-FFN	ShanghaiTech (82.6%), Avenue (91.7%)	Running, falling down, robbing, fighting.
[38], 2021	CNN, LSTM	Human skeleton, ConvLSTM	Weizmann (73.1%), KTH (93.4%), private (86.5%)	Punching, kicking
[39], 2021	RNN	Human skeleton, LSTM, GRU	UR Fall Detection and Fall Detection (98.2%)	Fall
[7], 2022	RNN	LSTM and GRU	Camera (84%)	Fall, fight
[13], 2022	RNN, CNN	3D-ConVNet, LSTM	RLVS (96.5%), Hockey (97%), violent flow (93.2%)	Violence
[20], 2022	CNN	Human skeleton, ConvLSTMs	Camera (85%)	Door blocking, door picking
[9], 2022	CNN	ConvLSTM	Abnormal Activities (97.64%)	Robbery, fight hijack, harassment
[40], 2022	CNN, LSTM	YOLOv5, ConvLSTM	Hockey fight (93.5%), Cigarette smoker (90%), Playing cards (93.8%)	Smoking, playing cards, fighting
[19], 2022	CNN	YOLOv5	Private (91%)	Not wearing safety helmet, entering dangerous area, smoking
[41], 2022	CNN, LSTM	ConvLSTM	Abnormal Activities (96.19%)	Begging, Drunkenness, Fight, Harassment, Hijack, Knife Hazard, Robbery, and Terrorism
[42], 2022	CNN	Human skeleton, YOLOv3, VGG16 pre-trained network	Camera (95%)	Walking, hugging, fighting

4.1. Two-Stream Convolutional Architecture (Dual-Stream CNNs)

Simonyan et al. [43] proposed a two-stream CNN, also known as a dual-stream CNN, to capture the spatial and temporal information, respectively. This model contains two networks (as shown in Figure 1) to capture the space and time information of video [5]. One network takes a single-frame image as the input, then obtains the spatial domain information by extracting the features hidden in the image. Whereas the input of the other network is a certain frame and n frames of images behind it in the video, which is responsible for processing the optical flow information in the video by stacking consecutive frames to extract temporal features. Finally, the outputs of the two networks are fused to obtain the classification result. Several researchers implemented two-stream CNN architectures for anomaly detection [43–46] and were shown to produce state-of-the-art results.

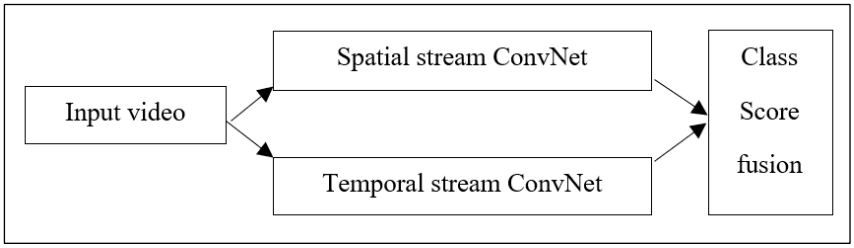

Figure 1. Two-stream network architecture [43].

4.2. 3D Convolution Architecture (3D-ConvNet)

In 2015, Tran et al. [47] suggested an approach for spatiotemporal feature extraction using deep 3-dimensional convolutional networks (3D-ConvNet) trained on a large-scale supervised video dataset. Furthermore, they demonstrated that the performance of convolutional 3D (C3D) features exceeded that of 2D ConvNet features on a wide range of video analysis tasks. By using 3D convolution and 3D pooling operations, 3D ConvNet can model both space and time to simultaneously find both spatial and temporal features in a video. In AD based on 3D-ConvNet, the ordinary convolution kernel is expanded to three dimensions, and the added dimension is responsible for processing information in the temporal domain. 3D convolution stacks multiple consecutive frames into a cube, and then uses the 3D convolution kernel in the composed cube to perform the operation [5]. The biggest advantage of the 3D-ConVNet structure is its speed, this encouraged many researchers to employ it in the AD area [8,12,18]. Figure 2 illustrates 3D operations. note that both input and filter have depth dimension D, and the 3D filter slides in the depth direction, 3D convolution operations output a volume. illustrates the 3D-convolution operation.

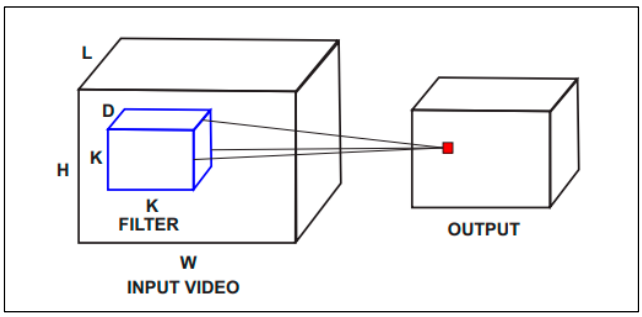

Figure 2. 3D-convolution operation.

4.3. ConvLSTM Architecture

ConvLSTM is CNN combined with an LSTM network. It is like LSTM, but convolutional operations are done during layer transitions [41]. ConvLSTM performs on

time-dependent data like video. As a result, the network is able to detect temporal and spatial correlations at the local level. ConvLSTM determines the future state of a particular cell in the grid using the inputs and past states of its local neighbors. By combining CNN and LSTM, AD is possible in multiple dimensions, including spatial, temporal, and any others that may be relevant to a given application [5]. Hence, correlating data from several dimensions allows for the detection of contextual anomaly structures that may not exhibit abnormal behavior in every single dimension. ConvLSTM-based AD methods are studied in [11,13,20,40,48]. They used a CNN to learn the space features in the input image, and then fed those features into an LSTM to identify features in the temporal domain. and then make a decision regarding whether or not each individual frame has displayed anomalous behavior. Figure 3 illustrates the architecture of ConvLSTM. Note that ConvLSTM layers are just like the LSTM, but internal matrix multiplications are exchanged with convolution operations. As a result, the data that flows through the ConvLSTM cells keeps the input dimension instead of being just a 1D vector with features.

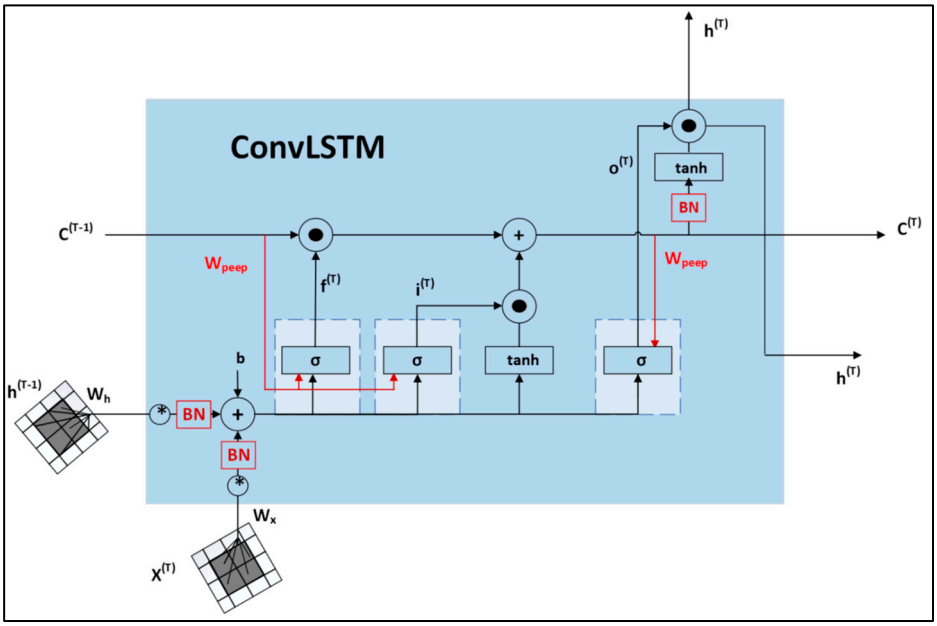

Figure 3. Structure of ConvLSTM [41].

4.4. Using Human Skeleton Data

Existing AD methods suffer from recognizing patterns in complex environments such as background variation, lighting changes, changes in pedestrian clothing, and a lack of dimensional information, all of which work against the efficacy of interactive behavior detection systems. The data on the human skeleton is a high level of abstraction from the body and can deal with interference pretty well [39]. Specifically, the methods based on human key points are used to detect the anomalies in the video because they can effectively eliminate background noise and extract human key points in crowded video scenes [13,38]. Multiple human skeleton-based methods have been proposed for action detection and recognition, such as Openpose, Mediapipe, and Alphapose. The studies in [20,49–51] used the Openpose method for human body extraction and recognition to provide a good basis for action detection in video. The Openpose algorithm is the first real-time solution for identifying key points in the human body, foot, hands, and face. It has also been added to the OpenCV library [52]. Figure 4 presents 18 joint points in the human body estimated by OpenPose. Mediapipe and Alphapose methods have been used in [10,53], respectively, to achieve the extraction and detection of key points on the human body. Mediapipe is an end-to-end, cross-platform skeletal software tool that works in real-time [53]. Furthermore, Microsoft Kinect sensor is one of the most widely used approaches to estimating a human's

3D position. This technique works by transforming 2D image detection from several camera views into 3D images [54,55]. Based on what has been found in the literature, a feature of the human skeleton could be a good way to recognize human behavior.

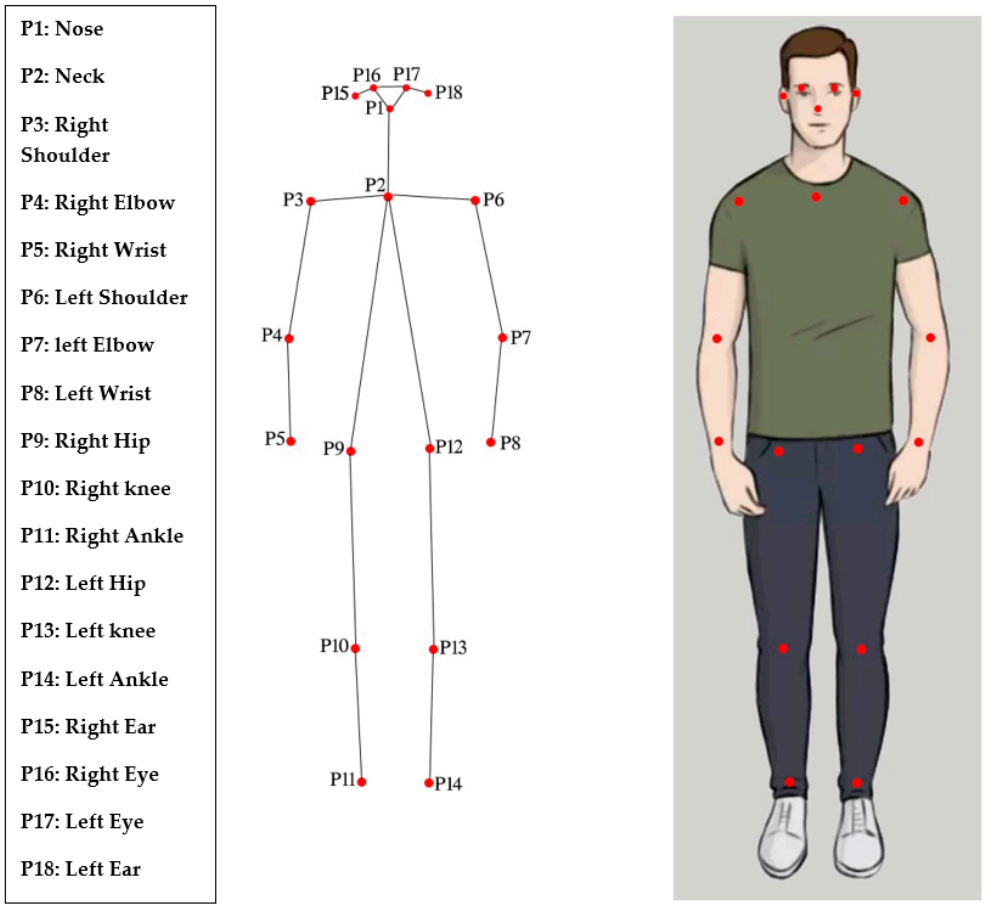

Figure 4. Key points of human bones.

4.5. Miscellaneous Architectures

This section, various AD architectures that have been shown to be effective and promising will be discussed. Ref. [21] proposed unsupervised learning approach based on a sequence of convolution auto-encoders and sequence-to-sequence LSTM (seq2seqLSTM) for the detection of anomalies in video. This work is implemented by passing a sequence of video frames to a convolution encoder to learn the spatial features from videos. A convolution encoder is trained by minimizing the loss between the model's output and the input. Thus, an encoded sequence of frames is obtained. Then, these encoded feature vectors of consecutive frames are fed to seq2seq LSTM to extract temporal features in the frames. Ref. [10] developed a framework of GRU layers and a dense Feed Forward Network (FFN) to estimate human activity. In this architecture, the output of one GRU unit is fed into the input of the next GRU unit. The generated feature formed by the GRU units is then passed to a fully connected layer, where it is mapped into 2D image coordinates. The fundamental benefits of this framework come from the use of dense FFN, which both ensures feature learning capability and takes advantage of the memorizing advantage provided by GRUs layers. Ref. [17] developed a combined method of CNN and RNN to classify human abnormalities. This method examines the face to spot anomalies like drug abuse, autism, and criminal behavior. It consists of convolution layers followed by the recurrent network. A CNN layer extracts the spatial features within the face regions of the image, while an RNN network takes into account the temporal dependencies that exist inside the image. Figure 5 illustrates the architecture of the combined CNN-RNN.

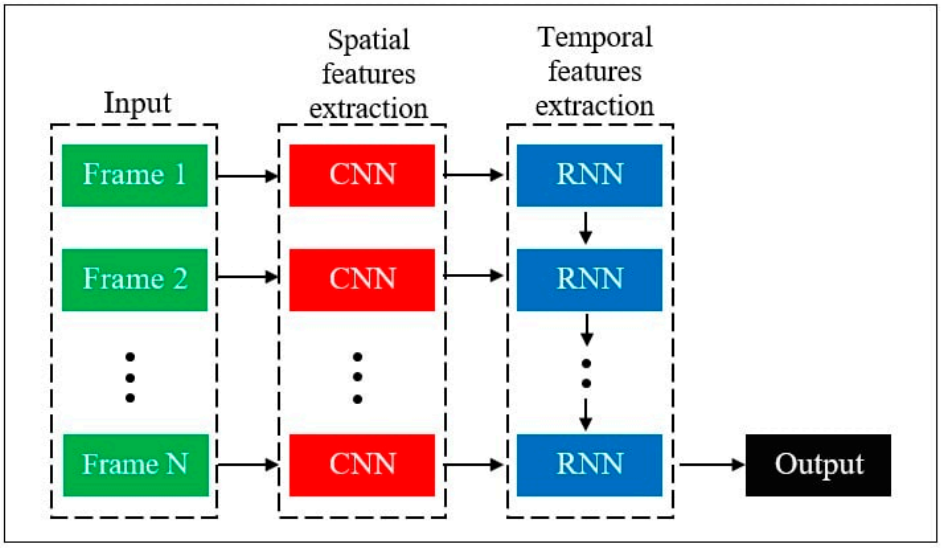

Figure 5. The combined CNN-RNN architecture.

5. Benchmark Datasets

Evaluation and comparing system performance is greatly aided by the benchmark dataset. Having a complete and reliable dataset allows us to evaluate the system's performance in a variety of ways. Many key factors must be taken into account when building a dataset, such as the availability of labeled data, activity type, size of samples, test environments, the diversity of the captured video, etc. Most researchers divide a dataset into two groups, training data and testing data, with certain percentages for each group, such as 70% and 30% [6,56] or 80% and 20% [11,17,38] from samples for training and testing, respectively. Few researchers divide their dataset into three sections: training, testing, and validation. Refs. [9,41] divided the dataset into 70% training, 20% testing, and 10% validation. Similarly, in ref. [40], the dataset is split into 80% training, 10% validation, and 10% testing. On the other hand, some researchers trained their models on a specific dataset and then tested them on a completely different dataset [57,58]. Table 2 lists the most widely used AD datasets that have been used to benchmark DL approaches in the academic literature. Furthermore, it provides the most relevant information needed while working with DL methods, such as the main reference, description, number of videos, examples of anomalies, and access details. A more detailed description of the existing datasets used for AD is presented in reference [59]. Figure 6 show examples of different anomalies in UCF-crime dataset.

Table 2. An overview of common datasets used for AD.

Dataset [Ref.]	Year	Description	No. of Videos	Resolution	Example Anomalies	URL
CAVIAR [60]	2004	It includes videos of two different situations. The sequences are ground truth labeled frame-by-frame with bounding boxes and a semantic description of the activity in each frame. There are 28 video sequences grouped into 6 different activity scenarios.	28	384 × 288	Fighting and leaving a package in a public place	https://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/ (accessed on 11 November 2022)

Table 2. *Cont.*

Dataset [Ref.]	Year	Description	No. of Videos	Resolution	Example Anomalies	URL
UMN [61]	2006	It's a collection of 11 videos depicting various escape scenarios across three indoor and outdoor scenes. Each clip starts with examples of normal behavior and then turns into abnormal examples.	11	320 × 240	People running (escape)	
UCSD-PED1 [62]	2010	It consists of clips of groups of people walking towards and away from the camera, and some amount of perspective distortion. There are 34 training and 36 testing videos, each containing 36 frames.	70	158 × 238	Movement of bikers, skaters, cyclist, small carts, people in a wheelchair	http://www.svcl.ucsd.edu/projects/anomaly/dataset.html (accessed on 11 November 2022)
UCSD-PED2 [62]	2010	It consists of a scene where most pedestrians move horizontally. The video footage of each scene is sliced into clips of 120–200 frames. There are 16 training videos and 12 testing ones.	28	240 × 360	Movement of bikers, skaters, cyclist, small carts, people in a wheelchair	http://www.svcl.ucsd.edu/projects/anomaly/dataset.html (accessed on 11 November 2022)
BEHAVE [63]	2010	It focuses on aberrant behavior associated with criminal activity. It has around 90,000 frames of humans identified by bounding boxes, with interacting groups classified into one of 6 different behaviors.	4	640 × 480	Chase, fight, and run	
Hockey fight [64]	2011	It is collected of hockey games and scenes from action movies to describe the violent behaviors in ice hockey matches. Each clip consisting of 50 frames, is manually labeled as "fight" or "non-fight"	1000	720 × 576	Fight	https://academictorrents.com/details/38d9ed996a5a75a039b84cf8a137be794e7cee89 (accessed on 15 November 2022)
HMDB-51 [56]	2011	It is collected from a variety of sources ranging from digitized movies to YouTube videos. In total, there are 51 action categories.	6766	Variable resolution	Shoot gun, climbing and falling	http://serre-lab.clps.brown.edu/resources/HMDB/ (accessed on 15 November 2022)
Violent Flow [65]	2012	Data is compiled from various sources to characterize the actions of crowds in public areas like parks, streets, and squares.	246	320 × 240	Violence	http://www.openu.ac.il/home/hassner/data/violentflows/ (accessed on 14 November 2022)
UCF-101 [66]	2012	A total of 27 h of footage, covering 101 different action categories, are included. Users uploaded videos with realistic camera movement and cluttered backgrounds to make the database.	13,320	320 × 240	Robbery, hijack, harassment, explosions, and fight	http://crcv.ucf.edu/data/UCF101.php (accessed on 14 November 2022)

Table 2. *Cont.*

Dataset [Ref.]	Year	Description	No. of Videos	Resolution	Example Anomalies	URL
CUHK Avenue [67]	2013	It contains 16 training and 21 testing video clips with total 30,652 frames which describe the movement and behavior of pedestrians, cars, cyclist.	37	640 × 360	Running, throwing objects, and loitering	http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html (accessed on 14 November 2022)
ActivityNet [68]	2015	It provides 203 activity classes, with an average of 137 videos per class, for a grand total of 849 video hours.	27,801	1280 × 720		http://www.activity-net.org (accessed on 15 November 2022)
Kinetics [69]	2017	It contains 400 human action classes, with 400–1150 clips for each action, each from a unique video. The clips average roughly 10 s in length and are all collected from various videos available on YouTube.	306,245	variable resolution	Violence	https://www.deepmind.com/open-source/kinetics (accessed on 16 November 2022)
ShanghaiTech Campus [70]	2017	It has 13 scenes with complex light conditions and camera angles. It contains 130 abnormal events and over 270, 000 training frames.	330	846 × 480	Brawling, chasing, skaters, bikers, and trolley on the pedestrian walkways	https://svip-lab.github.io/dataset/campus_dataset.html (accessed on 16 November 2022)
UCF-Crime [71]	2018	It has 128 h of 1900 long and untrimmed real-world surveillance videos, with 13 realistic anomalies as well as normal activities	1900	variable resolution	Abuse, Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, Vandalism.	http://crcv.ucf.edu/projects/real-world/ (accessed on 16 November 2022)
RLVS [72]	2019	It consists of violent clips that involve fights in many different environments, such as the street, jails, and schools. The nonviolent videos also feature human activities, including playing sports, exercising, and eating.	2000	Average size of 397 × 511	Fight and Violence	https://www.kaggle.com/datasets/mohamedmustafa/real-life-violence-situations-dataset (accessed on 17 November 2022)

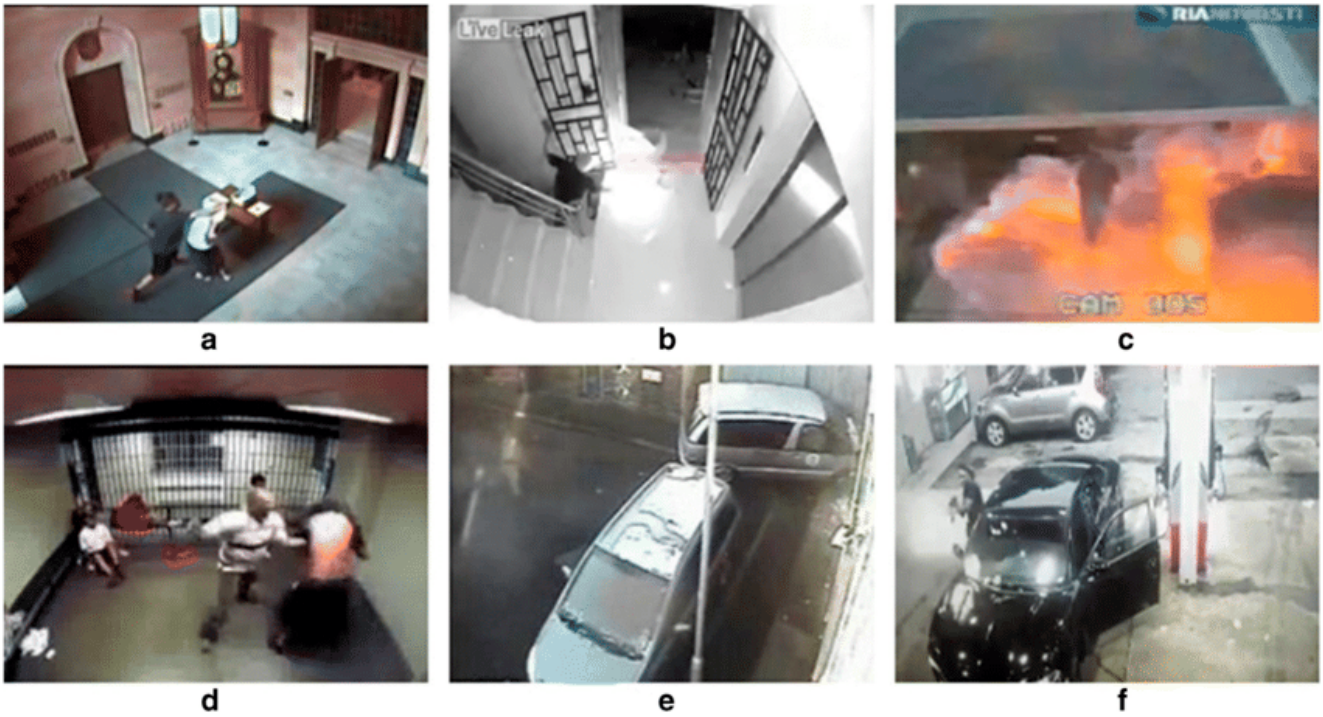

Figure 6. Examples of different anomalies in UCF-crime dataset (**a**) Abuse (**b**) Arson (**c**) Explosion (**d**) Fight (**e**) Road Accident (**f**) Shooting.

6. Anomaly Detection Approach Performance Metrics

The effectiveness of AD systems has been evaluated in many ways by researchers. AD models aim to achieve low false positive (FP) and false negative (FN) rates. On another side, True positive (TP) and true negative (TN) rates should also be high. How many negative (normal) and positive (anomaly) examples are correctly labeled is denoted by TN and TP, respectively. While the FP and FN counts indicate how many instances were incorrectly labeled as positive or negative. Table 3 presents the most significant metrics used for evaluating AD model performance [27,73–76].

Table 3. Evaluation metrics used for AD approaches.

Metric	Definition	Equation
Accuracy	It measures the number of anomalous and normal instances that are successfully classified with respect to the overall dataset. Accuracy can be a useful measure if we have a similar balance in the dataset.	$\frac{TP+TN}{TP+TN+FP+FN}$
Equal Error rate (EER)	It is a metric that evaluates the proportion of anomalies and normal instances that are misclassified with respect to the overall dataset. It's used to show biometric performance.	$\frac{FP+FN}{TP+TN+FP+FN}$
Recall (Sensitivity) (True Positive Rate)	The ratio of detected anomalies to total anomalies is calculated. Recall is very used when you have to correctly classify some event that has already occurred.	$\frac{TP}{TP+FN}$
Precision (Detection rate)	It is a metric that compares the number of real anomalies discovered to the total number of anomalies. It calculates the accuracy of the True Positive.	$\frac{TP}{TP+FP}$
Specificity (True Negative Rate)	It determines the percentage of the samples that were correctly labeled as normal. specificity is important when the objective is to minimize the number of negative examples that are incorrectly classified.	$\frac{TN}{FP+TN}$

Table 3. *Cont.*

Metric	Definition	Equation
False Positive Rate (FPR)	It is the ratio of the number of anomalous instances that are incorrectly classified in relation to all normal instances.	$\frac{FP}{(FP+TN)}$
False Negative Rate (FNR)	It measures the ratio of normal instances that are incorrectly classified in relation to all normal instances.	$\frac{FN}{(FP+TN)}$
F1-Score	It calculates the harmonic Mean between recall and precision rates. The greater the F1-Score, the better is the performance of the model. It's often used when class distribution is uneven.	$2 \times \frac{Precision \times Recall}{Precision+Recall}$
J Score	It is a single statistic that captures the performance of a binary classification test.	Sensitivity + Specificity − 1
Percentage of Wrong Classifications (PWC)	It calculates the ratio between the number of incorrect predictions and the total number of predictions.	$100 \times \frac{FP+FN}{TP+TN+FP+FN}$
Receiver operating characteristic curve (ROC)	It gives details on a curve that represents the percentage of anomalies that were correctly recognized against those that were missed at varying thresholds.	
Area under ROC curve (AUC)	It is the area under the curve of the plot of FPR vs. TPR at different points in [0, 1]. As the value increases, our model's accuracy improves. It yields good results when the observations are balanced between each class.	$\frac{s_p - n_p(n_n+1)/2}{n_p \times n_n}$

TP: true positive, TN: true negative, FP: false positive, FN: false negative, s_p: sum of all positive ranked samples, n_n and n_p: number of negative and positive samples, respectively.

7. Applications of AD for Video

Here, some applications of DL-based AD in videos will be discussed. More applications and techniques used are illustrated in Table 4.

Table 4. Examples of applications of DL-based AD.

Application Type	Technique Used	Ref.
Automated surveillance	CNN	[77]
	CNN and LSTM	[78]
Autonomous driving	Autoencoder + semantic segmentation	[79]
	GAN + Post hoc statistics	[80]
	CNN + Gaussian Processes	[81]
Industrial automation	LSTM and autoencoder	[82]
	LSTM, CNN, autoencoder	[83]
Intelligent traffic monitoring	YOLOv5 and decision tree	[84]
Surgical Robotics	Deep Residual Autoencoder	[85]

7.1. Autonomous Driving

Autonomous vehicles mainly depend on techniques of perceptual vision that use intelligent algorithms. One of the crucial challenges in this domain is to handle unexpected situations and detect anomaly actions in time to avoid accident, like a person suddenly passing the street, ghost driver. The ability to reliably detect such anomalies is essential for improving automated driving safety, since it can significantly reduce the incidents involving autonomous vehicles [86,87].

7.2. Automated Surveillance

Intelligent surveillance video, commonly known as closed-circuit television (CCTV), has been widely applied in many public places, such as roads [88], hospitals [89], banks [90], campuses [40], elevators [20], and private homes [91]. Such intelligent surveillance systems provide numerous advantages in lifestyle including protection of human resources, financial burden reduction, and anomalies behavior detection in time with high accuracy. The fact that there is no clear definition of an anomaly is a big problem that hurts how well these systems work.

7.3. Industrial Automation

Due to the spread of smart factory services, AI-based research is being conducted to predict and diagnose manufacturing facility breakdowns or manufacturing site efficiency. However, because of the characteristics of manufacturing data, such as a severe class imbalance of abnormalities and ambiguous label information that distinguishes abnormalities, developing classification or AD models is highly difficult. Industrial visual defect detection is hard because mistakes can be small, like thin scratches, or big, like missing parts [92].

7.4. Medical AD

The medical analysis depends on the diagnosis task, which in turn is related to AD in the physiological data of patients since it captures the unique features in the physiological data of patients. Thus, the identification of anomalies in medical data is considered a sensitive task in such a field [93]. AD systems in medical tests face extra challenges because they are directly related to human life and health. Further, there are many patient-specific characteristics that should be taken into account when designing these systems, such as age and gender, that lead to variations in data samples. For these reasons, supervised learning algorithms are mostly used when developing models of medical AD due to their high ability to distinguish between normal and abnormal samples [94].

8. Research Challenges in DL-Based AD Approaches

In this section, we highlight the major issues to be resolved in current AD approaches that prevent them from being applied effectively to detect anomalous actions in videos.

8.1. Anomaly Characteristics

The task of AD from a video is a challenging task because the anomaly activities occur for a short duration of time and have a low probability of occurrence, as well as the context-dependent nature of anomalies, and one anomaly class may have completely different features from other anomaly classes. Furthermore, a diversity of anomaly scenarios [10,21].

8.2. Anomaly Definition

Current concept of normal or abnormal behavior may not be adequately representational in the future [16], and the definition of abnormality itself may change over time. The lack of a clear definition of an anomaly in video surveillance is a big problem that makes it hard for AD systems to work.

8.3. Environmental Factors

Environmental challenges indicate large variations in camera viewpoint and motion, cluttered background, and foreground scale variation [95]. Furthermore, changes in position, human occlusion, low-quality and noisy video, illumination changes, weather conditions, and appearances of the actors [8,56]. Some researchers tried to address these issues. For example, ref. [23] used a panoramic camera to achieve 360-degree video acquisition in order to address the problems of camera viewpoint and person occlusion. Similarly, ref. [15] proposed a hybrid approach of distributed and centralized processing to detect abnormal behaviors of various target entities. On the other hand, human skeleton data is a high level of abstraction from the body and it has been utilized to address the

problems of sparsely or moderately crowded video scenes and complex environments such as background variation, lighting changes, and noisy video [13,38].

8.4. Division of Dataset

To guarantee a consistent comparative analysis of the presented DL models, it is necessary to suggest a strong mechanism for dataset splitting. For DL networks to learn the optimal model parameters, there is a certain amount of data needed for training. Although there are AD datasets out there, the majority of them do not clearly separate the training and testing data. Furthermore, most DL models have been optimized for a single video or a group of similar videos, so the training and testing data are highly similar [57,96]. Moreover, the model would have an unfair advantage compared to other models if it were trained using some frames from the test videos. In addition, the documentation of incomparable outcomes is hampered by the fact that different articles use different data split methodologies. One possible solution to address this issue is testing the model over completely unseen videos to show its robustness. However, there is a need to have well-defined data division strategies for both training and testing.

8.5. Data Diversity

It is challenging to gather large and diverse datasets for training AD networks [74]. Additionally, the application of DL-based models is hindered by the lack of anomalous ground truth data, the ambiguous nature of anomalies, and the data imbalance between normal and anomalous samples [97,98]. Transfer learning approaches can be a valuable approach to overcome the limited training data problem. Furthermore, data augmentation techniques can be utilized to artificially expand the size of the training dataset.

8.6. Data Annotation

Training a DL network requires large-scale labeled datasets. Additionally, human motion time-series data annotation is a cumbersome and long process. Moreover, data with ambiguous definitions, like abnormal human behavior, is challenging to annotate since it is hard to understand which human action includes the anomaly [99]. There is a need to employ scientific methods to detect, fix, and reduce the errors that occur in data annotation in order to ensure that the final deliverable data is of the highest possible quality, consistency, and integrity.

8.7. Feature Normalization

Human behavior data needs to be normalized in a way that can be processed in real time, hence it is important to establish effective ways for doing so. Feature normalization can help find and recognize human behavior when data from many different people is used to choose the right set of features to improve the overall system performance and computational complexity.

8.8. Model Generalization

Since human behavior is so context-and environment-dependent, it is challenging to design a single model to detect anomalies across all scenarios and domains. One possible solution to address this issue is training the model on data sets from many different people and scenarios.

8.9. Real Time Systems

A majority of the current approaches utilized for video AD are very time-and space-consuming. Because of this, real-world applications cannot make use of these techniques [95]. Thus, it is important to provide algorithms for real-time AD. Possible solutions include using online learning techniques to analyze anomalies in real-time by combining online learning with deep models and making deeper networks with more layers and fewer neurons [26].

8.10. Lightweight Models

DL networks require a lot of computing power to perform video AD efficiently. So, it is important to propose lightweight models that can work well with limited resource devices like mobile devices. Some possible solutions have been proposed by [100–102].

9. Future Directions

9.1. Aerial Surveillance

Activity recognition and AD in aerial videos are considered important domains of research due to their contribution to many vital tasks such as search and rescue and aerial surveillance. Unmanned aerial vehicles, commonly known as "drones," are usually used to capture aerial video data. Some challenges make it hard to find anomalies in aerial videos, like when the drone is moving in the opposite direction from the object or when the object and drone are moving at different speeds [103,104]. To overcome these difficulties, researchers in the future will need to develop innovative algorithms specifically for use with aerial videos.

9.2. AD from Moving Cameras

The majority of existing research for DL-based AD deals with captured data from fixed cameras, while AD from moving cameras is still a limited domain. Thus, this field is a promising area of study [105]. Future work will require more effort to be put into the collection of data and the development of algorithms for AD in moving cameras.

9.3. Self-Supervised Learning in Video

The activity detection techniques can offer solutions for the free labels to develop self-supervised systems for several systems, like object detection [106], video order prediction [107], and video representation learning [108]. Future work needs to create self-supervised algorithms by leveraging existing techniques like intelligent surveillance and AD techniques.

9.4. Human–Robot Collaboration

The human–robot collaboration will play a crucial role in the future of industrial production lines [5]. Future trends in DL-based human behavior recognition might be toward human movement prediction by understanding human intentions. There is a need to create new artificial algorithms that use the patterns of human movement as input data to modify the robot's trajectory or speed to avoid people. This research domain will improve the social environment of humans and robots.

9.5. Ensemble Approaches

Ensemble methods are a promising research domain. Ensemble approaches have been shown to be effective in improving the efficiency of AD [109]. The ensemble detection of deviations is another promising area for future study because it has the ability to greatly improve the algorithms' detection accuracy. However, this field of research is still undiscovered and requires more comprehensive study.

10. Conclusions

This review aimed to be a significant research contribution to the study of DL in the intelligent surveillance domain by analyzing and summarizing the DL techniques utilized in AD for video streaming. In particular, our broad study used two categories to categorize AD. The first category considered the number of frames used during detection, while the second one the number of anomalies in a scene. Moreover, our study analyzed the efficacy of many popular DL techniques for detecting anomalies, categorizing them according to the network type and architectural design. Moreover, the benchmark datasets and performance metrics used to evaluate the effectiveness of DL approaches were listed in detail. Furthermore, our contribution highlighted the applications as well as the key issues

in DL-based AD approaches that are still open and need to be addressed for efficient AD. Finally, we are convinced that the community researchers working on this topic will surely find this review helpful in gaining a better understanding of this crucial area of research. Then, our main goal was to encourage researchers to carry out more research in this area so that it can move forward in the near future.

Author Contributions: Conceptualization, S.A.J., K.A.H. and H.K.H.; methodology, S.A.J. and L.A.; software, S.A.J. and L.A.; validation, S.A.J., K.A.H., H.K.H., L.A. and J.S.; data curation, S.A.J., K.A.H., H.K.H., L.A. and J.S.; writing—original draft preparation, S.A.J. and L.A.; writing—review and editing, S.A.J., K.A.H., H.K.H., L.A. and J.S.; project administration, S.A.J., K.A.H., H.K.H., L.A. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: Laith Alzubaidi would like to acknowledge the support received through the following funding schemes of Australian Government: ARC Industrial Transformation Training Centre (ITTC) for Joint Biomechanics under grant IC190100020.

Data Availability Statement: All relevant dataset links were provided in the main review paper content.

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

References

1. Dávila-Montero, S.; Dana-Lê, J.A.; Bente, G.; Hall, A.T.; Mason, A.J. Review and Challenges of Technologies for Real-Time Human Behavior Monitoring. *IEEE Trans. Biomed. Circuits Syst.* **2021**, *15*, 2–28. [CrossRef] [PubMed]
2. 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.
3. Ruff, L.; Kauffmann, J.R.; Vandermeulen, R.A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T.G.; Müller, K.-R. A Unifying Review of Deep and Shallow Anomaly Detection. *Proc. IEEE* **2021**, *109*, 756–795. [CrossRef]
4. Al-Dhamari, A.; Sudirman, R.; Mahmood, N.H. Transfer Deep Learning along with Binary Support Vector Machine for Abnormal Behavior Detection. *IEEE Access* **2020**, *8*, 61085–61095. [CrossRef]
5. Yuan, J.; Wu, X.; Yuan, S. A Rapid Recognition Method for Pedestrian Abnormal Behavior. In Proceedings of the 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 10–12 July 2020; pp. 241–245.
6. Bian, C.; Wang, L.; Gu, H.; Zhou, F. Abnormal Behavior Recognition Based on Edge Feature and 3D Convolutional Neural Network. In Proceedings of the 2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Zhanjiang, China, 16–18 October 2020; pp. 1–6.
7. Gorodnichev, M.G.; Gromov, M.D.; Polyantseva, K.A.; Moseva, M.S. Research and Development of a System for Determining Abnormal Human Behavior by Video Image Based on Deepstream Technology. In Proceedings of the 2022 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), Sankt Petersburg, Russia, 31 May–4 June 2022; pp. 1–9.
8. Cao, B.; Xia, H.; Liu, Z. A Video Abnormal Behavior Recognition Algorithm Based on Deep Learning. In Proceedings of the 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 18–20 June 2021; Volume 4, pp. 755–759.
9. Vrskova, R.; Hudec, R.; Kamencay, P.; Sykora, P. Recognition of Human Activity and Abnormal Behavior Using Deep Neural Network. In Proceedings of the 2022 14th International Conference Elektro, Krakow, Poland, 23–26 May 2022; pp. 1–4.
10. Fan, B.; Li, P.; Jin, S.; Wang, Z. Anomaly Detection Based on Pose Estimation and GRU-FFN. In Proceedings of the 2021 IEEE Sustainable Power and Energy Conference (iSPEC), Nanjing, China, 23–25 December 2021; pp. 3821–3825.
11. Traoré, A.; Akhloufi, M.A. Violence Detection in Videos Using Deep Recurrent and Convolutional Neural Networks. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada,, 11–14 December 2020; pp. 154–159.
12. Emad, M.; Ishack, M.; Ahmed, M.; Osama, M.; Salah, M.; Khoriba, G. Early-Anomaly Prediction in Surveillance Cameras for Security Applications. In Proceedings of the 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt,, 26–27 May 2021; pp. 124–128.
13. Chexia, Z.; Tan, Z.; Wu, D.; Ning, J.; Zhang, B. A Generalized Model for Crowd Violence Detection Focusing on Human Contour and Dynamic Features. In Proceedings of the 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Taormina, Italy, 16–19 May 2022; pp. 327–335.
14. Zhang, W.; Miao, Z.; Xu, W. A Video Anomalous Behavior Detection Method Based on Multi-Task Learning. In Proceedings of the 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 15–17 April 2022; pp. 396–400.
15. Alkanat, T.; Groot, H.G.J.; Zwemer, M.; Bondarev, E.; de Peter, H.N. Towards Scalable Abnormal Behavior Detection in Automated Surveillance. In Proceedings of the 2021 4th International Conference on Artificial Intelligence for Industries (AI4I), Laguna Hills, CA, USA, 20–22 September 2021; pp. 21–24.

16. Tang, X.; Astle, Y.S.; Freeman, C. Deep Anomaly Detection with Ensemble-Based Active Learning. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 1663–1670.
17. Kabir, M.M.; Safir, F.B.; Shahen, S.; Maua, J.; Binte Awlad, I.A.; Mridha, M.F. Human Abnormality Classification Using Combined CNN-RNN Approach. In Proceedings of the HONET 2020—IEEE 17th International Conference on Smart Communities: Improving Quality of Life using ICT, IoT and AI, Charlotte, NC, USA, 14–16 December 2020; pp. 204–208. [CrossRef]
18. Heo, T.; Nam, W.; Paek, J.; Ko, J. Autonomous Reckless Driving Detection Using Deep Learning on Embedded GPUs. In Proceedings of the 2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), Delhi, India, 10–13 December 2020; pp. 464–472.
19. Xiao, Y.; Wang, Y.; Li, W.; Sun, M.; Shen, X.; Luo, Z. Monitoring the Abnormal Human Behaviors in Substations Based on Probabilistic Behaviours Prediction and YOLO-V5. In Proceedings of the 2022 7th Asia Conference on Power and Electrical Engineering (ACPEE), Hangzhou, China, 15–17 April 2022; pp. 943–948.
20. Shi, Y.; Guo, B.; Xu, Y.; Xu, Z.; Huang, J.; Lu, J.; Yao, D. Recognition of Abnormal Human Behavior in Elevators Based on CNN. In Proceedings of the 2021 26th International Conference on Automation and Computing (ICAC), Portsmouth, UK, 2–4 September 2021; pp. 1–6.
21. Pawar, K.; Attar, V. Application of Deep Learning for Crowd Anomaly Detection from Surveillance Videos. In Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 28–29 January 2021; pp. 506–511.
22. Wang, Z.; Jiang, K.; Hou, Y.; Dou, W.; Zhang, C.; Huang, Z.; Guo, Y. A Survey on Human Behavior Recognition Using Channel State Information. *IEEE Access* **2019**, *7*, 155986–156024. [CrossRef]
23. Li, J.; Xie, H.; Zang, Z.; Wang, G. Real-Time Abnormal Behavior Recognition and Monitoring System Based on Panoramic Video. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 7129–7134.
24. Marsiano, A.F.D.; Soesanti, I.; Ardiyanto, I. Deep Learning-Based Anomaly Detection on Surveillance Videos: Recent Advances. In Proceedings of the 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Yogyakarta, Indonesia, 20–21 September 2019; pp. 1–6.
25. Chalapathy, R.; Chawla, S. Deep Learning for Anomaly Detection: A Survey. *arXiv* **2019**, arXiv:1901.03407.
26. Pawar, K.; Attar, V. Deep Learning Approaches for Video-Based Anomalous Activity Detection. *World Wide Web* **2019**, *22*, 571–601. [CrossRef]
27. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* **2021**, *8*, 1–74. [CrossRef]
28. 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]
29. Alzubaidi, L.; Al-Shamma, O.; Fadhel, M.A.; Farhan, L.; Zhang, J.; Duan, Y. Optimizing the Performance of Breast Cancer Classification by Employing the Same Domain Transfer Learning from Hybrid Deep Convolutional Neural Network Model. *Electronics* **2020**, *9*, 445. [CrossRef]
30. Ali, L.R.; Jebur, S.A.; Jahefer, M.M.; Shaker, B.N. Employing Transfer Learning for Diagnosing COVID-19 Disease. *Int. J. Onl. Eng.* **2022**, *18*, 31–42. [CrossRef]
31. Alzubaidi, L.; Al-Amidie, M.; Al-Asadi, A.; Humaidi, A.J.; Al-Shamma, O.; Fadhel, M.A.; Zhang, J.; Santamaría, J.; Duan, Y. Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data. *Cancers* **2021**, *13*, 1590. [CrossRef] [PubMed]
32. Alzubaidi, L.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J.; Santamaría, J.; Duan, Y.; Oleiwi, S.R. Towards a Better Understanding of Transfer Learning for Medical Imaging: A Case Study. *Appl. Sci.* **2020**, *10*, 4523. [CrossRef]
33. Liu, Y.; Li, Z.; Zhou, C.; Jiang, Y.; Sun, J.; Wang, M.; He, X. Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Trans. Knowl. Data Eng.* **2019**, *32*, 1517–1528. [CrossRef]
34. Pimentel, T.; Monteiro, M.; Veloso, A.; Ziviani, N. Deep Active Learning for Anomaly Detection. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
35. Pang, G.; van den Hengel, A.; Shen, C.; Cao, L. Deep Reinforcement Learning for Unknown Anomaly Detection. *arXiv* **2020**, arXiv:2009.06847.
36. Aberkane, S.; Elarbi, M. Deep Reinforcement Learning for Real-World Anomaly Detection in Surveillance Videos. In Proceedings of the 2019 6th International Conference on Image and Signal Processing and their Applications (ISPA), Mostaganem, Algeria, 24–25 November 2019; pp. 1–5.
37. 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.
38. Naik, A.J.; Gopalakrishna, M.T. Deep-Violence: Individual Person Violent Activity Detection in Video. *Multimed. Tools Appl.* **2021**, *80*, 18365–18380. [CrossRef]
39. Lin, C.-B.; Dong, Z.; Kuan, W.-K.; Huang, Y.-F. A Framework for Fall Detection Based on Openpose Skeleton and Lstm/Gru Models. *Appl. Sci.* **2020**, *11*, 329. [CrossRef]
40. Khayrat, A.; Malak, P.; Victor, M.; Ahmed, S.; Metawie, H.; Saber, V.; Elshalakani, M. An Intelligent Surveillance System for Detecting Abnormal Behaviors on Campus Using YOLO and CNN-LSTM Networks. In Proceedings of the 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), Cairo, Egypt, 8–9 May 2022; pp. 104–109.

41. Vrskova, R.; Hudec, R.; Kamencay, P.; Sykora, P. A New Approach for Abnormal Human Activities Recognition Based on ConvLSTM Architecture. *Sensors* **2022**, *22*, 2946. [CrossRef]
42. Ali, M.A.; Hussain, A.J.; Sadiq, A.T. Deep Learning Algorithms for Human Fighting Action Recognition. *Int. J. Online Biomed. Eng.* **2022**, *18*, 71–87.
43. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–11.
44. Huang, X.; He, P.; Rangarajan, A.; Ranka, S. Intelligent Intersection: Two-Stream Convolutional Networks for Real-Time near-Accident Detection in Traffic Video. *ACM Trans. Spat. Algorithms Syst.* **2020**, *6*, 1–28. [CrossRef]
45. Hao, W.; Zhang, R.; Li, S.; Li, J.; Li, F.; Zhao, S.; Zhang, W. Anomaly Event Detection in Security Surveillance Using Two-Stream Based Model. *Secur. Commun. Netw.* **2020**, *2020*, 8876056. [CrossRef]
46. Jamadandi, A.; Kotturshettar, S.; Mudenagudi, U. Two Stream Convolutional Neural Networks for Anomaly Detection in Surveillance Videos. In *Smart Computing Paradigms: New Progresses and Challenges*; Springer: Singapore, 2020; pp. 41–48.
47. 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.
48. Abdali, A.-M.R.; Al-Tuma, R.F. Robust Real-Time Violence Detection in Video Using Cnn and Lstm. In Proceedings of the 2019 2nd Scientific Conference of Computer Sciences (SCCS), Baghdad, Iraq, 27–28 March 2019; pp. 104–108.
49. Lin, F.-C.; Ngo, H.-H.; Dow, C.-R.; Lam, K.-H.; Le, H.L. Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection. *Sensors* **2021**, *21*, 5314. [CrossRef] [PubMed]
50. Li, S.; Yi, J.; Farha, Y.A.; Gall, J. Pose Refinement Graph Convolutional Network for Skeleton-Based Action Recognition. *IEEE Robot. Autom. Lett.* **2021**, *6*, 1028–1035. [CrossRef]
51. Ali, M.A.; Hussain, A.J.; Sadiq, A.T. Human Fall Down Recognition Using Coordinates Key Points Skeleton. *Int. J. Online Biomed. Eng.* **2022**, *18*, 88–104.
52. Lathifah, N.; Lin, H.-I. A Brief Review on Behavior Recognition Based on Key Points of Human Skeleton and Eye Gaze To Prevent Human Error. In Proceedings of the 2022 13th Asian Control Conference (ASCC), Jeju Island, Republic of Korea, 4–7 May 2022; pp. 1396–1403.
53. Zhang, F.; Bazarevsky, V.; Vakunov, A.; Tkachenka, A.; Sung, G.; Chang, C.-L.; Grundmann, M. Mediapipe Hands: On-Device Real-Time Hand Tracking. *arXiv* **2020**, arXiv:2006.10214.
54. Jia, J.-G.; Zhou, Y.-F.; Hao, X.-W.; Li, F.; Desrosiers, C.; Zhang, C.-M. Two-Stream Temporal Convolutional Networks for Skeleton-Based Human Action Recognition. *J. Comput. Sci. Technol.* **2020**, *35*, 538–550. [CrossRef]
55. Agahian, S.; Negin, F.; Köse, C. An Efficient Human Action Recognition Framework with Pose-Based Spatiotemporal Features. *Eng. Sci. Technol. Int. J.* **2020**, *23*, 196–203. [CrossRef]
56. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A Large Video Database for Human Motion Recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
57. Patil, P.W.; Murala, S. MSFgNet: A Novel Compact End-to-End Deep Network for Moving Object Detection. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 4066–4077. [CrossRef]
58. Patil, P.W.; Biradar, K.M.; Dudhane, A.; Murala, S. An End-to-End Edge Aggregation Network for Moving Object Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8149–8158.
59. Patil, N.; Biswas, P.K. A Survey of Video Datasets for Anomaly Detection in Automated Surveillance. In Proceedings of the 2016 Sixth International Symposium on Embedded Computing and System Design (ISED), Patna, India, 15–17 December 2016; pp. 43–48.
60. Fisher, R.B. The PETS04 Surveillance Ground-Truth Data Sets. In Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Prague, Czech Republic, 10 May 2004; pp. 1–5.
61. Mehran, R.; Oyama, A.; Shah, M. Abnormal Crowd Behavior Detection Using Social Force Model. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 935–942. [CrossRef]
62. 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; pp. 1975–1981.
63. Blunsden, S.; Fisher, R.B. The BEHAVE Video Dataset: Ground Truthed Video for Multi-Person Behavior Classification. *Ann. BMVA* **2010**, *4*, 4.
64. Bermejo Nievas, E.; Deniz Suarez, O.; Bueno García, G.; Sukthankar, R. Violence Detection in Video Using Computer Vision Techniques. In *Computer Analysis of Images and Patterns*; Springer: Cham, Switzerland, 2011; pp. 332–339.
65. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent Flows: Real-Time Detection of Violent Crowd Behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 1–6.
66. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv* **2021**, arXiv:1212.0402.
67. 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.

68. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A Large-Scale Video Benchmark for Human Activity Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
69. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P. The Kinetics Human Action Video Dataset. *arXiv* **2017**, arXiv:1705.06950.
70. 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.
71. 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.
72. Soliman, M.M.; Kamal, M.H.; Nashed, M.A.E.-M.; Mostafa, Y.M.; Chawky, B.S.; Khattab, D. Violence Recognition from Videos Using Deep Learning Techniques. In Proceedings of the 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 8–10 December 2019; pp. 80–85.
73. Mandal, M.; Vipparthi, S.K. An Empirical Review of Deep Learning Frameworks for Change Detection: Model Design, Experimental Frameworks, Challenges and Research Needs. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6101–6122. [CrossRef]
74. Lindemann, B.; Maschler, B.; Sahlab, N.; Weyrich, M. A Survey on Anomaly Detection for Technical Systems Using LSTM Networks. *Comput. Ind.* **2021**, *131*, 103498. [CrossRef]
75. Kalsotra, R.; Arora, S. A Comprehensive Survey of Video Datasets for Background Subtraction. *IEEE Access* **2019**, *7*, 59143–59171. [CrossRef]
76. Wu, P.; Liu, J.; Shen, F. A Deep One-Class Neural Network for Anomalous Event Detection in Complex Scenes. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 2609–2622. [CrossRef]
77. Zhao, Y.; Man, K.L.; Smith, J.; Guan, S.-U. A Novel Two-Stream Structure for Video Anomaly Detection in Smart City Management. *J. Supercomput.* **2022**, *78*, 3940–3954. [CrossRef]
78. 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. *Futur. Gener. Comput. Syst.* **2022**, *129*, 286–297. [CrossRef]
79. Ohgushi, T.; Horiguchi, K.; Yamanaka, M. Road Obstacle Detection Method Based on an Autoencoder with Semantic Segmentation. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
80. Nitsch, J.; Itkina, M.; Senanayake, R.; Nieto, J.; Schmidt, M.; Siegwart, R.; Kochenderfer, M.J.; Cadena, C. Out-of-Distribution Detection for Automotive Perception. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, Indianapolis, IN, USA, 19–22 September 2021; Volume 2021, pp. 2938–2943. [CrossRef]
81. Ryan, C.; Murphy, F.; Mullins, M. End-to-End Autonomous Driving Risk Analysis: A Behavioural Anomaly Detection Approach. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1650–1662. [CrossRef]
82. Lindemann, B.; Fesenmayr, F.; Jazdi, N.; Weyrich, M. Anomaly Detection in Discrete Manufacturing Using Self-Learning Approaches. *Procedia CIRP* **2019**, *79*, 313–318. [CrossRef]
83. Maschler, B.; Knodel, T.; Weyrich, M. Towards Deep Industrial Transfer Learning for Anomaly Detection on Time Series Data. In Proceedings of the 2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Vasteras, Sweden, 7–10 September 2021; pp. 1–8.
84. Aboah, A. A Vision-Based System for Traffic Anomaly Detection Using Deep Learning and Decision Trees. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4207–4212.
85. Samuel, D.J.; Cuzzolin, F. Unsupervised Anomaly Detection for a Smart Autonomous Robotic Assistant Surgeon (SARAS) Using a Deep Residual Autoencoder. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7256–7261. [CrossRef]
86. Breitenstein, J.; Termöhlen, J.-A.; Lipinski, D.; Fingscheidt, T. Corner Cases for Visual Perception in Automated Driving: Some Guidance on Detection Approaches. *arXiv* **2021**, arXiv:2102.05897.
87. Ferreira, R.S.; Guérin, J.; Guiochet, J.; Waeselynck, H. SiMOOD: Evolutionary Testing Simulation with Out-Of-Distribution Images. In Proceedings of the 27th IEEE Pacific Rim International Symposium on Dependable Computing (PRDC 2022), Beijing, China, 22 November–1 December 2022.
88. Siddique, A.; Afanasyev, I. Deep Learning-Based Trajectory Estimation of Vehicles in Crowded and Crossroad Scenarios. In Proceedings of the 2021 28th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 27–29 January 2021; pp. 413–423.
89. Prati, A.; Shan, C.; Wang, K.I.-K. Sensors, Vision and Networks: From Video Surveillance to Activity Recognition and Health Monitoring. *J. Ambient Intell. Smart Environ.* **2019**, *11*, 5–22.
90. Bakunah, R.A.; Baneamoon, S.M. A Hybrid Technique for Intelligent Bank Security System Based on Blink Gesture Recognition. *J. Phys. Conf. Ser.* **2021**, *1962*, 12001. [CrossRef]
91. Rego, A.; Ramírez, P.L.G.; Jimenez, J.M.; Lloret, J. Artificial Intelligent System for Multimedia Services in Smart Home Environments. *Cluster Comput.* **2022**, *25*, 2085–2105. [CrossRef]
92. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards Total Recall in Industrial Anomaly Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 14318–14328.

93. Fernando, T.; Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Deep Learning for Medical Anomaly Detection–A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–37. [CrossRef]
94. Fernando, T.; Denman, S.; Ahmedt-Aristizabal, D.; Sridharan, S.; Laurens, K.R.; Johnston, P.; Fookes, C. Neural Memory Plasticity for Medical Anomaly Detection. *Neural Netw.* **2020**, *127*, 67–81. [CrossRef]
95. Xu, K.; Jiang, X.; Sun, T. Anomaly Detection Based on Stacked Sparse Coding with Intraframe Classification Strategy. *IEEE Trans. Multimed.* **2018**, *20*, 1062–1074. [CrossRef]
96. Akilan, T.; Wu, Q.J.; Safaei, A.; Huo, J.; Yang, Y. A 3D CNN-LSTM-Based Image-to-Image Foreground Segmentation. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 959–971. [CrossRef]
97. Maschler, B.; Weyrich, M. Deep Transfer Learning for Industrial Automation: A Review and Discussion of New Techniques for Data-Driven Machine Learning. *IEEE Ind. Electron. Mag.* **2021**, *15*, 65–75. [CrossRef]
98. Vu, H.; Phung, D.; Nguyen, T.D.; Trevors, A.; Venkatesh, S. Energy-Based Models for Video Anomaly Detection. *arXiv* **2017**, arXiv:1708.05211.
99. Miki, D.; Chen, S.; Demachi, K. Unnatural Human Motion Detection Using Weakly Supervised Deep Neural Network. In Proceedings of the 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), Irvine, CA, USA, 21–23 September 2020; pp. 10–13.
100. Mehmood, A. LightAnomalyNet: A Lightweight Framework for Efficient Abnormal Behavior Detection. *Sensors* **2021**, *21*, 8501. [CrossRef] [PubMed]
101. Osifeko, M.O.; Hancke, G.P.; Abu-Mahfouz, A.M. SurveilNet: A Lightweight Anomaly Detection System for Cooperative IoT Surveillance Networks. *IEEE Sens. J.* **2021**, *21*, 25293–25306. [CrossRef]
102. Chang, S.; Li, Y.; Shen, S.; Feng, J.; Zhou, Z. Contrastive Attention for Video Anomaly Detection. *IEEE Trans. Multimed.* **2021**, *24*, 4067–4076. [CrossRef]
103. Mandal, M.; Kumar, L.K.; Vipparthi, S.K. Mor-Uav: A Benchmark Dataset and Baselines for Moving Object Recognition in Uav Videos. In Proceedings of the 28th ACM International Conference on Multimedia, New York, NY, USA, 12–16 October 2020; pp. 2626–2635.
104. Chen, X.; Li, Z.; Yang, Y.; Qi, L.; Ke, R. High-Resolution Vehicle Trajectory Extraction and Denoising from Aerial Videos. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 3190–3202. [CrossRef]
105. Jiang, C.; Paudel, D.P.; Fofi, D.; Fougerolle, Y.; Demonceaux, C. Moving Object Detection by 3d Flow Field Analysis. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1950–1963. [CrossRef]
106. Fang, Z.; Jain, A.; Sarch, G.; Harley, A.W.; Fragkiadaki, K. Move to See Better: Self-Improving Embodied Object Detection. *arXiv* **2020**, arXiv:2012.00057.
107. Xu, D.; Xiao, J.; Zhao, Z.; Shao, J.; Xie, D.; Zhuang, Y. Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10334–10343.
108. Han, T.; Xie, W.; Zisserman, A. Video Representation Learning by Dense Predictive Coding. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
109. Al-amri, R.; Murugesan, R.K.; Man, M.; Abdulateef, A.F.; Al-Sharafi, M.A.; Alkahtani, A.A. A Review of Machine Learning and Deep Learning Techniques for Anomaly Detection in IoT Data. *Appl. Sci.* **2021**, *11*, 5320. [CrossRef]