

Review

Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey

Huu-Thanh Duong , Viet-Tuan Le  and Vinh Truong Hoang * 

Faculty of Information Technology, Ho Chi Minh City Open University, 97 Vo Van Tan, District 3, Ho Chi Minh City 700000, Vietnam; thanh.dh@ou.edu.vn (H.-T.D.); tuan.lv@ou.edu.vn (V.-T.L.)

* Correspondence: vinh.th@ou.edu.vn

Abstract: Anomaly detection in video surveillance is a highly developed subject that is attracting increased attention from the research community. There is great demand for intelligent systems with the capacity to automatically detect anomalous events in streaming videos. Due to this, a wide variety of approaches have been proposed to build an effective model that would ensure public security. There has been a variety of surveys of anomaly detection, such as of network anomaly detection, financial fraud detection, human behavioral analysis, and many more. Deep learning has been successfully applied to many aspects of computer vision. In particular, the strong growth of generative models means that these are the main techniques used in the proposed methods. This paper aims to provide a comprehensive review of the deep learning-based techniques used in the field of video anomaly detection. Specifically, deep learning-based approaches have been categorized into different methods by their objectives and learning metrics. Additionally, preprocessing and feature engineering techniques are discussed thoroughly for the vision-based domain. This paper also describes the benchmark databases used in training and detecting abnormal human behavior. Finally, the common challenges in video surveillance are discussed, to offer some possible solutions and directions for future research.

Keywords: abnormal human activity recognition; deep learning; video surveillance; anomaly detection



Citation: Duong, H.-T.; Le, V.-T.; Hoang, V.T. Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey. *Sensors* **2023**, *23*, 5024. <https://doi.org/10.3390/s23115024>

Academic Editors: Bo Zhou and Sungho Suh

Received: 27 February 2023

Revised: 28 April 2023

Accepted: 8 May 2023

Published: 24 May 2023



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

1. Introduction

With today's increasing demand for security, especially in public places such as airports, train stations, supermarkets, schools, and crowded street, surveillance cameras are used for monitoring daily activities and detecting abnormal events. This task focuses on the localization of anomalies using both temporal and partial information in videos. Anomalies can be defined as events deviating from normal behavior [1], e.g., fighting, sneaking, or unattended bags at an airport. The purpose of using surveillance cameras is the early detection of anomalous human behaviors. This is a critical task in many cases where human intervention is necessary, e.g., for crime prevention or countering terrorism. However, this process requires labor-intensive and continuous human attention, which is a tedious process, since abnormal events only happen 0.01% of the time and 99.9% of the surveillance time is wasted [2]. Moreover, a surveillance system produces a lot of redundant video data, which require unnecessary storage space. For reducing human errors and storage costs, it is necessary to build an efficient surveillance system for detecting any strange behaviors that may lead to dangerous situations. This requires deep and comprehensive study of human activity recognition, to understand the features representative of each action.

Anomaly detection in video has a wide range of applications, such as for traffic accident detection, criminal activity detection, and illegal activity detection. In addition, detecting anomalous items or abandoned objects, such as guns or knives, is necessary in sensitive areas.

Anomaly detection has become an active area of research in recent years. With the purpose of trying to conduct an automated process for detecting abnormal events, many

methods have been proposed. The essential idea is to first learn normal patterns from training videos, then extract representations of normal cases. If any events deviate from these representations, an abnormal event can be detected. However, there are still a lot challenges in video surveillance and human activity recognition. These difficulties are related to the feature extraction stage, where occlusion, overlapping, cluttered backgrounds, sensor noise, low illumination, and dynamic background changes may impact a system's performance [3]. It is quite challenging to extract robust and discriminative features for training models due to these various challenges. Moreover, anomaly detection is also dependent on the scene context [4], where one action is considered abnormal in one scene but may be normal in another. This requires a vast amount of data for training, to verify all cases happening in the real world.

This survey explores the current deep learning methods that have been used in video surveillance. In particular, it focuses on the understanding of abnormal human behaviors, since the identification and recognition of actions is key for an intelligent video system. This survey's scope encompasses vision-based approaches, since this is becoming a popular trend in action recognition. The significant contributions of this survey are as follows:

- A comprehensive review of vision-based human activity recognition for video surveillance.
- A description of popular databases used in anomaly detection.
- Analysis of data processing and feature engineering for deep learning models.
- Discussion of recent deep learning models, along with their advantages and disadvantages.
- Identifying existing challenges and future research for anomaly detection in video surveillance.

In order to prepare for the survey, an essential step was to search and filter recent papers on video surveillance using deep learning techniques for understanding human activities. The research papers were download using relevant index terms, such as *video surveillance*, *deep learning activity recognition*, *human action representation*, *abnormal behavior detection*. The references from these papers were extracted, validated, and added to the survey. After that, research relevant to video surveillance was also included. Finally, the general structure and representative methods were provided in detail, to give an insight into each learning model.

The organization of this survey is as follows: In Section 2, the background and the related surveys are provided. A discussion of common databases used in video surveillance is provided in Section 3. Section 4 explains data processing and feature engineering in the vision-based domain. A detailed analysis of the state-of-the-art deep learning models is provided in Section 5. Section 6 discusses the existing challenges in anomaly detection and gives directions for future research. Finally, Section 7 concludes the survey.

2. Background Knowledge and Related Works

2.1. Background Knowledge

Human action recognition (HAR) is a fundamental problem in computer vision and has been studied for a long time. The purpose of HAR is to identify the action taking place in a video, in order to understand and produce an analysis of specific events. A video's spatial and temporal information plays an important role in HAR, to correctly identify human actions and classify the video. The detection of abnormal behavior in video surveillance is basically used to ensure security in both indoor and outdoor locations, such as in airports and train stations. Abnormal human action recognition (AbHAR) can be considered a particular problem in HAR. The problem with AbHAR in videos is that it can vary widely, and there is no single approach to solving all problem cases. The common approaches to AbHAR rely on feature extraction from image sequences. These features are built for the tasks of object detection, pose estimation, and finding the dense trajectories that are useful in HAR.

AbHAR is used in video surveillance for monitoring behavior and people's activities, with the purpose of ensuring security and giving instructions. This requires a certain general background in many video processing domains. Video surveillance includes the implementation of knowledge of feature extraction, scene understanding, object tracking, object identification, and model generation. Specifically, the feature extraction step needs to be addressed carefully, since it has a great impact on the whole system. Traditional machine learning algorithms have achieved remarkable results in AbHAR, based on learning shallow features from video data. Methods such as random forest (RF) [5], Bayesian networks [6], Markov models [7], and support vector machine (SVM) [8] have been used to understand and recognize human behavior. These methods are heavily dependent on preprocessing and handcrafted features, which require a lot of time and resources to process. Furthermore, they do not scale well for different datasets and show a poor performance with real cases [9]. In recent years, deep learning methods have achieved great interest from the research community, since they can automatically extract learning features and have shown promising results for difficult research topics, such as object detection and recognition [10], image classification [11], and natural language processing (NLP) [12]. Compared to traditional ML methods, deep learning is a multistage learning process that automatically extracts the representative features for a specific task utilizing several hidden layers [13]. These features are called deep features, and they can be scaled reasonably well to various scenarios. Deep learning has recently been applied to HAR and AbHAR, and has proven to be very efficient for video surveillance systems [14].

Figure 1 shows the diagram of the primary processing steps for AbHAR in video surveillance. In general, there are four main steps: object segmentation, object classification, object tracking, and action recognition. Object segmentation localizes moving objects (mainly human targets) on the scene. Object classification identifies the type of each object of interest for later processing. These two steps are heavily reliant on extracting features from video data. Then the targets of interest are tracked through the frame sequence using object tracking algorithms. Object tracking methods refer to motion estimation and re-identification for the tracking of targets. Lastly, action recognition learns representative features to classify and understand which type of anomalous activities represent each tracking object. If abnormal behavior is detected, then an alarm or notification is sent to the authorities.

2.2. Related Works

Anomaly detection in video surveillance has been studied for decades. The topic has attracted much attention from the research community, due to its varied applications. A vast amount of methods have been proposed for both traditional and deep learning approaches to AbHAR. In order to keep track of all available studies, thorough surveys are of considerable interest. In 1997 and 1999, there were reviews [15,16] on human motion analysis applied in HAR. In 2005, automated video surveillance systems were surveyed in [17]. Three years later, a study of behavior analysis for homeland security applications [18] was published. In 2012, the authors in [19] provided a survey of anomaly detection in video surveillance applications, within different contexts.

In 2013, a survey was performed of vision-based HAR [20], which reviewed the existing research on HAR. The survey focused on the use of the self-evaluation method to extract important features and also suggested some directions for future research. In 2014, the authors in [21] summarized the different techniques used for intelligent surveillance systems in public places. Another review of crowded scene analysis [22] was conducted in 2014. In the subsequent two years, the authors in [23] focused on various HAR methods for video streams. One year later, there was another comprehensive review [24] covering both handcrafted and deep feature representations. At the same time, behavior representation and behavior modeling were both discussed in [25].

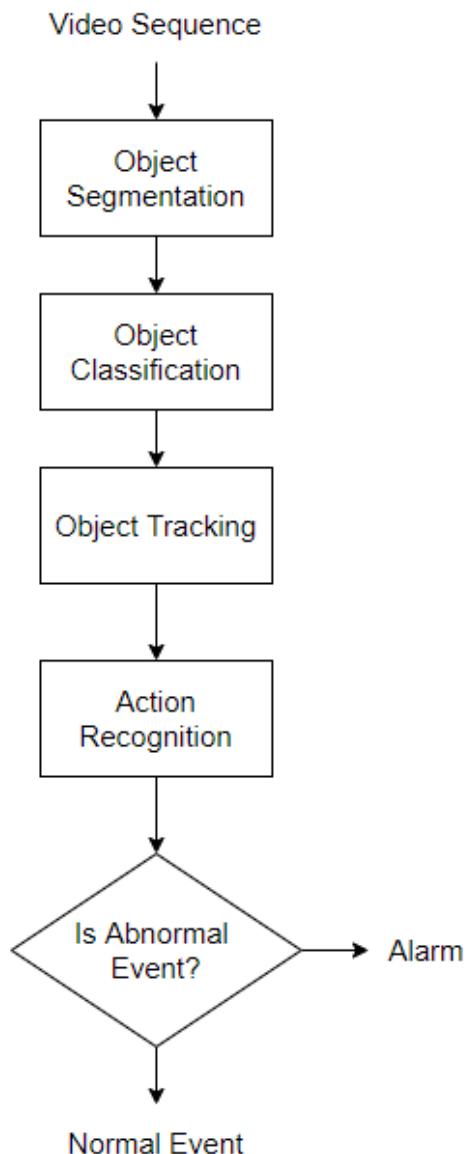


Figure 1. General Process of Video Surveillance.

A survey [26] in 2018 covered the use of RGB-D-based methods for action detection and classification. The authors also provided a detailed discussion of the strengths and weakness of each anomaly detection method. In the same year, the relationship between HAR and data stream mining was thoroughly summarized in [27]. There was also a discussion of unsupervised and semi-supervised anomaly detection in [28]. In 2019, there were a lot of surveys of HAR. The authors of [29] performed a comprehensive review focused on the three different aspects of feature presentation, interaction recognition, and human action detection methods. The authors of [30] reported different processing techniques for video surveillance. The review in [31] focused on both handcrafted and deep approaches for different types of two-dimensional and three-dimensional data. Most recently, the survey in [32] detailed various techniques for single-scene video anomaly detection. The authors in [33] provided a comprehensive survey that discussed both sensor-based and vision-based human activity recognition.

A survey of network anomaly detection [34] was performed by Ahmed et al. They focused on four types of network attack; namely, denial of service, probe, user to root, and remote to user. They grouped the research methods into four categories, including classification-based network anomaly detection, statistical anomaly detection, information

theory, and clustering-based. They also discussed some intrusion detection datasets. Another survey [35] reviewed anomaly based network intrusion detection systems (NIDSs), which is an important field in cybersecurity. The authors grouped methods into six categories, as in [34], and provided further details about benchmark datasets. A more detailed survey [36] on network anomaly detection was published in 2019. This survey was split into two main parts. The first part discussed anomaly detection background analysis, considering traffic anomalies, network data types, and intrusion detection systems. Network traffic anomalies were divided into two categories: anomaly categorization based on its nature, and based on its causal aspect. They also discussed two types of intrusion detection systems (IDS), which are IDS types using a monitored platform and IDS types using a detection technique. The second part mainly discussed anomaly detection methods, techniques, and systems. This part included statistical, clustering, finite state machine, classification, information theory, evolutionary computation, and hybrid/other methods.

3. Benchmark Databases

Anomaly detection has been extensively studied in many fields, especially in computer vision, for the purpose of learning and understanding activity recognition. This task can become very challenging, due to the complexity of real-world situations. There are infinite abnormal events, and it is impossible to gather all kinds. Fortunately, many databases have been developed to help the scientist and researcher with this task. These databases were mainly acquired in the visible spectrum, so they are sensitive to occlusion and illumination changes. In this section, we briefly introduce some popular databases that researchers are currently using for detecting behavioral anomalies. We arrange the databases in chronological order, from the oldest to the newest database. For each database, we summarize information of the release year, type of database (i.e., single-scene or multiscene), details of the sensor used (RGB or thermal, resolution, FPS), description of anomalous activities, and sample images.

3.1. CASIA Action Database

The CASIA Action database [37] was released in 2007 and is a type of RGB single-scene dataset, i.e., it was recorded with color sensors in a certain location. This database includes human outdoor activities captured from different viewing angles. There are eight types of action for a single person, such as walk, run, bend, jump, crouch, faint, wander, and punching a car, from 24 different subjects. There are also seven types of two person interaction, such as rob, fight, follow, follow and gather, meet and part, meet and gather, overtake. The database includes two main parts: single person actions, and interactions (see Figure 2). A description is given below:

- Single person action:
 - Walk: One subject walking along the road.
 - Run: One subject running along the road.
 - Bend: One subject bending his/her body while walking.
 - Jump: One subject jumping along the road.
 - Crouch: One subject crouching once while walking along the road.
 - Faint: One subject falling down on the ground while walking along the road.
 - Wander: One subject wandering around.
 - Punching a car: One subject punching a car.
- Interaction:
 - Rob: One subject robbing another one.
 - Fight: Two subjects fighting with each other.
 - Follow: One subject following another till the end.
 - Follow and gather: One subject following another and then walking together with the other to the end.
 - Meet and part: Two subjects meeting each other and then departing.

- Meet and gather: Two subjects meeting each other and then walking together till the end.
- Overtake: One subject overtaking another.

In order to capture the videos, three noncalibrated cameras were used simultaneously from different viewing angles, which were a horizontal view, angled view, and top-down view. These videos were encoded at a frame rate of 25 FPS and were compressed with the huffyuv codec in avi format. The original spatial resolution of the video frames was reduced to 320×240 . Each video lasts from 5 s to 30 s, due to the different kind of action. Figure 2 shows some images from the database.

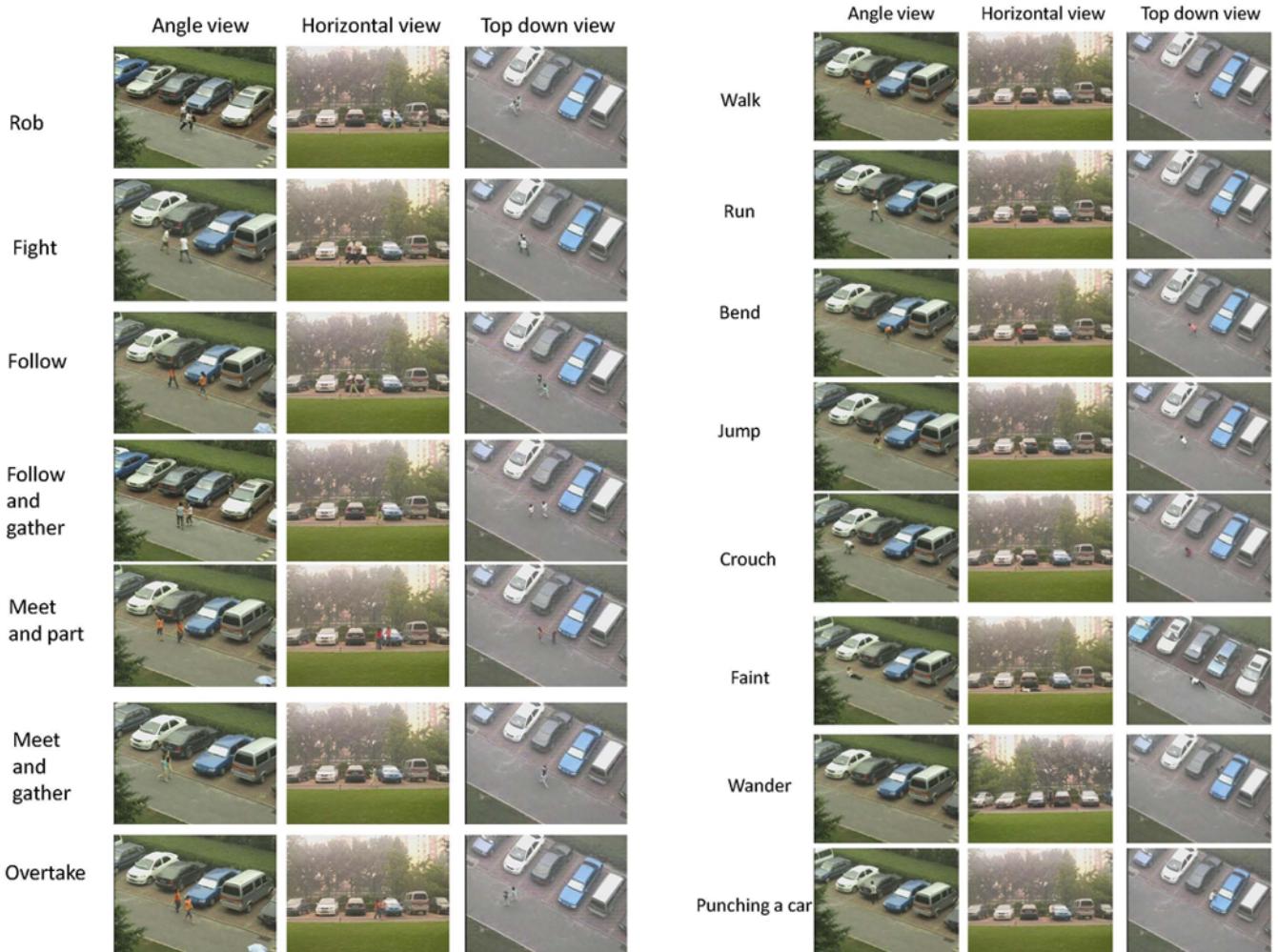


Figure 2. Sample images from the CASIA database. From left to right: single person action, two person interaction [37].

3.2. Subway Database

The subway database was introduced by Adam et al. in their 2008 paper [38]. This is also a type of single-scene database. This database comprises two long video recordings monitoring people at a subway entrance and exit, see Figure 3. There is no spatial ground truth available. The video was recorded in grayscale format at 15 FPS with a resolution of 512×384 , and it has 125,475 frames in total. The anomalous events are mainly wrong directions, loitering, no payment, people jumping or squeezing through turnstiles, and a janitor cleaning the walls.



Figure 3. Sample images from the Subway Database. From left to right: subway entrance, subway exit [38].

3.3. UMN Crowd Abnormality Database

The UMN database was released in 2009 and first described in the paper titled “Abnormal crowd behavior detection using social force model” by R. Mehran et al. [39]. The video scenario simulated a populated area, where actors are wandering around a certain location and run away in escape mode with abnormal behavior. Therefore, this database can be considered a multiscene database, meaning that it was recorded in different locations.

The database has a total of 11 short videos, which are aggregated into one long video of 4 min 17 s with 7739 frames. The small videos begin with normal behavior, then change to abnormal. There is one scenario for the indoor scene and two outdoors. All the videos have the same frame rate of 30 FPS and were recorded at a resolution of 640×480 using a static camera. The ground truth is a temporal annotation. Figure 4 provides some sample images from the video.



Figure 4. Sample images from the UMN database [39].

3.4. Anomalous Behavior Database

The Anomalous Behavior Database [40] was released in 2010 by York University. The database comprises eight videos (multiscene) recorded in various challenging conditions, such as illumination effects, scene clutter, variable target appearance, rapid motion, and camera jitter. The database was also provided with a spatiotemporal ground truth, along with software for detecting abnormal events in certain parts of each video. The image sequences in this database mainly focus on the activities of humans and vehicles in certain public locations, such as an airport, river, sea, and on a train, see Figure 5:

- Traffic–Train: This video records daily activity on a train. This is a very challenging video, since the lighting conditions change drastically and there is camera jitter. The video includes 19,218 frames of RGB image with a resolution of 288×386 and a frame rate of 25 FPS. The anomalous event is the movement of a passenger.
- Bellevue: This video includes cars moving through an intersection. Video is recorded in grayscale format at a resolution of 320×240 , with a frame rate of 10 FPS, and it has 2918 frames in total. The anomalous event is cars entering the thoroughfare from the left or right.

- Boat–Sea: This video describes a passing boat as an abnormal event. The video was recorded in RGB format at a resolution of 720×576 , with a frame rate of 19 FPS, and it has 450 frames in total.
- Boat–River: This video illustrates a boat passing on a river as an abnormal event. The video was recorded in RGB format at a resolution of 720×576 , with a frame rate of 5 FPS, and it has 250 frames in total.
- Canoe: This video describes a canoe passing on a river as an abnormal event. The video was recorded in RGB format at a resolution of 320×240 , with a frame rate of 30 FPS, and it has 1050 frames in total.
- Camouflage: This video illustrates a person walking in camouflage. The right motion is learned as the normal behavior, and the opposite is the abnormal behavior. The video was recorded in RGB format at a resolution of 320×240 , with a frame rate of 30 FPS, and it has 1629 frames in total.
- Airport–WrongDir: This video records people walking in a line at an airport. The video was recorded in RGB format at a resolution of 300×300 , with a frame rate of 25 FPS, and it has 2200 frames in total. The anomalous event is people moving in the wrong direction.

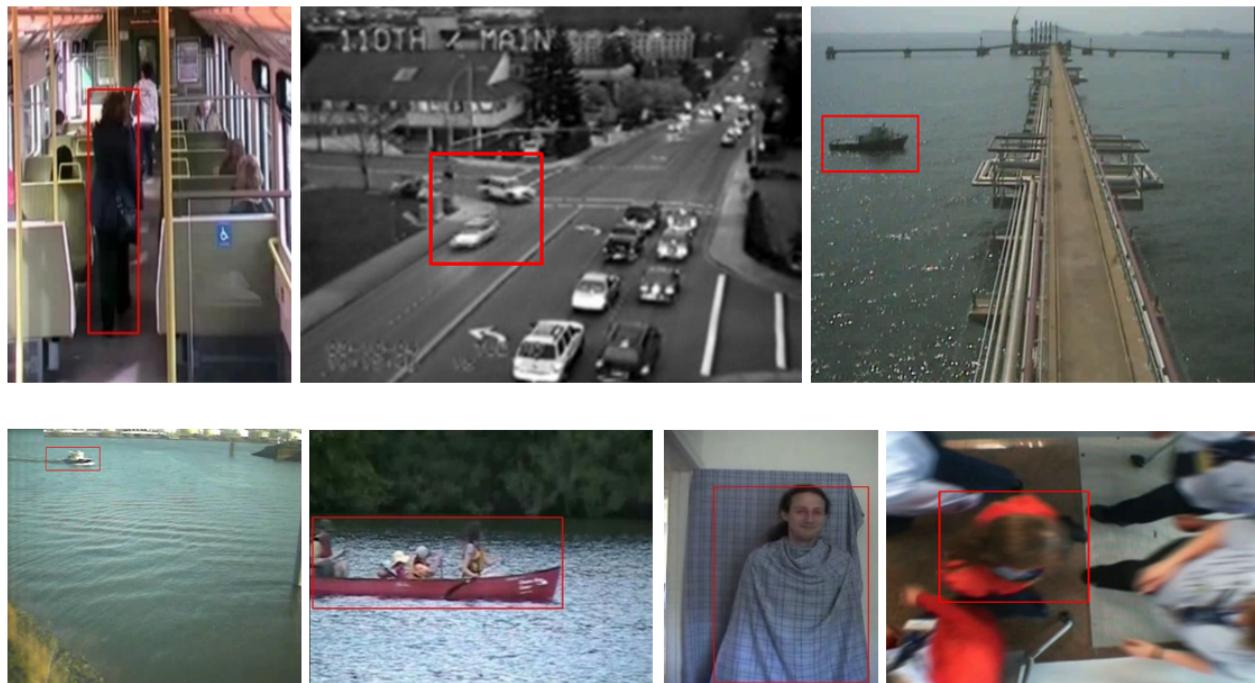


Figure 5. Sample images from the Anomalous Behavior Database. From left to right, and top to bottom: Traffic–Train, Traffic–Bellevue, Boat–Sea, Boat–River, Canoe, Camouflage, Airport–WrongDir [40].

3.5. Avenue Database

This database [41] was released in 2013 and contains 37 videos, divided into 16 normal videos for training and 21 abnormal videos for testing. The database is of a RGB single-scene type. There is a total of 47 abnormal events, categorized into three main subjects: strange actions, wrong direction, and abnormal object. These videos were captured at CUHK campus avenue, with 30,652 frames (15,328 training, 15,324 testing) in total. Each image sequence has a resolution of 640×360 and frame rate of 25 FPS. The author provided both temporal and spatial annotations. Three main types of abnormal events (see Figure 6) are defined below:

- Strange actions: behaviors such as running, throwing objects, and loitering.
- Wrong direction: people moving in the wrong direction.

- Abnormal objects: people carrying some strange objects with them, such as a bicycle.



Figure 6. Sample images from the Avenue Database. From left to right: running, throwing object, and loitering [41].

3.6. UCSD Anomaly Detection Database

This database [42] was released in 2013 and includes two subdatasets: Pedestrian 1, and Pedestrian 2. Both include a grayscale sequence of images record at 10 FPS with a resolution of 238×158 for pedestrian 1, and 360×240 for pedestrian 2. Each dataset is a single-scene dataset. Both have training videos containing only normal behaviors and testing videos containing abnormal events. A static camera was used to record the dataset and was setup on an elevator overlooking the pedestrian walkways. The abnormal events include the following:

- The circulation of non-pedestrian entities on the walkways such as bikers, skaters, and small carts.
- Anomalous pedestrian motion patterns, such as people walking across a walkway or in the grass that surrounds it.

The Pedestrian 1 (Ped1) dataset includes 34 normal training videos and 36 abnormal testing videos of groups of people walking towards and away from the camera. These abnormal cases are mainly related to abnormal vehicles, such as bicycles and cars entering the crowd. Some sample images are shown in Figure 7.



Figure 7. Ped1 subdataset with nonpedestrian entity [42].

The Pedestrian 2 (Ped2) dataset contains 16 training videos and 12 testing videos with 12 abnormal events. Ped2 includes scenes with pedestrian movement parallel to the camera plane. The definition of an anomaly for Ped2 is the same as for Ped1. Some example images from the Ped2 dataset are shown in Figure 8.



Figure 8. Ped2 subdataset with nonpedestrian entity [42].

3.7. ShanghaiTech Campus Database

This database [43] was released in 2016. It contains 330 training videos with only normal events, and 107 testing videos with 130 abnormal events. The total frames are 317,398, with 17,090 irregularity frames. The database was acquired using an RGB camera with a resolution of 856×480 at 24 FPS, overlooking pedestrian walkways. It consists of 13 scenes (multiscene) with complex light conditions, camera angles, and various anomaly types, mainly related to strange objects, wrong direction, and strange actions (see Figure 9).

- Strange actions: behaviors such as running, robbing, pushing, jumping, jumping over the fence, dropping, throwing objects, and fighting. Below are some sample images.
- Wrong direction: there is the case where people usually follow a normal direction, but someone does the inverse.
- Abnormal objects: this is a case where a person carries a strange object with them, such as a bicycle or baby stroller.



Figure 9. Sample images from the Avenue Database. From left to right: robbing, throwing object, and fighting [43].

3.8. UCF-Crime Database

The UCF-Crime database was released in 2018 [2]. This database is a compilation of 128 h of 1900 internet videos, taken from many RGB cameras at different locations (multiscene). The anomalous events include abuse, arrest, arson, assault, road accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism. These videos cover 13 real-world situations and can be used for two primary tasks: the event recognition of 13 group activities, and anomaly detection in each specific group. The authors only provided temporal annotations. Figure 10 shows some sample images from the database.



Figure 10. Sample images from the UCF-Crime Database. From left to right: abuse, fighting, and shoplifting [2].

3.9. Street Scene Database

The Street Scene database is the most recent database, being released in 2020 [44]. This dataset is a RGB single-scene type. There are a total of 203,257 image sequences, extracted from the original videos at a frame rate of 15 FPS. The database consists of 205 anomalous events, such as jaywalking, biker outside lane, and loitering... recorded at a resolution of 1280×720 , see Figure 11. This database is challenging, due to the various activities, occlusions, shadows, and moving backgrounds.

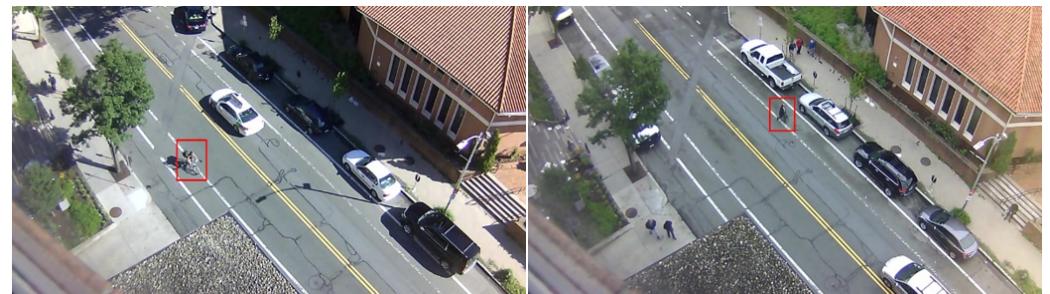


Figure 11. Sample images from the Street Scene Database. From left to right: biker outside lane and jaywalking [44].

4. Preprocessing Data for Video Anomaly Detection

The discussed benchmark databases are just collections of raw data of videos and images. These data need to be preprocessed before feeding into the ML algorithms. Data preparation and feature engineering are considered two important processes that greatly affect the performance of a video surveillance system. For the vision-based domain, these processes include several steps, such as background construction, foreground extraction, and feature representation extraction. The main purpose of the feature engineering process is reducing noise, selecting important representation features, transforming high-dimension features into the sub-space domain without losing valuable information, and reducing the overfitting problem. However, there are a lot of challenges, such as variation in lighting conditions, cluttered backgrounds, occlusion, or spurious interactions between subjects.

4.1. Segmentation

The first step in data preprocessing is segmentation. Segmentation is used to extract the target subjects from images or videos. Segmentation includes background construction and foreground extraction techniques. Background construction algorithms model background information by trying to determine a scene's global representative features, and the identified subjects are then analyzed based on the difference between the current frame and the constructed background [45]. Background construction techniques are used for tracking of fast objects in a scene, and they are computer efficient when used with fixed-cameras. Some statistical methods of background construction, such as [46–48], can work well in the multi-modal domain with adaptive parameters; however, their performance is greatly reduced by environmental noise or poor lighting conditions. The neural network based-techniques reported in [45,49,50] can overcome these limitations, but they easily overfit

data. The choice of technique mainly depends on the application's purpose. For dynamic background recording with moving cameras, it is necessary to use a foreground extraction-based segmentation algorithm. Both spatial and temporal information are analyzed from a video sequence to extract target subjects from the background. The authors in [51,52] used an optical flow technique to deal with occlusion and distortion, and successfully extracted the targets from the video recording by moving the camera; however, their methods are complex and time-consuming. Temporal information was also used in [53,54] to perform foreground segmentation, which was sensitive to noise but required a low computing power. References [55,56] used Markov random fields to preserve the boundaries and handle complex backgrounds, but these methods are not computer efficient.

Many of the methods proposed for video anomaly detection have tried to exploit both background and foreground information. Lai et al. [57] introduced a network that included two decoders to generate a future frame and RGB difference. Doshi et al. [58] used a pretrained object detection model (YOLO) to capture location and appearance features by detecting objects in a video. On the other hand, Cai et al. [59] used an image clip and its optical flow clip as input to capture structure and motion information. Then, two decoders generated a future frame and optical flow image using the fused feature, which was extracted from two encoders.

4.2. Feature Extraction and Selection

Handcrafted-feature-based extraction [60,61] was used to extract useful features for understanding human behaviors. These methods are restricted to certain conditions, and lack the flexibility to adapt to new environments. They are also time consuming and computer inefficient. Researchers turned to new representative features, which are categorized into three different types: local, global, and semantic features. These features have clearly shown their advantages and robustness to noise and dynamic environments.

- Local representative features use local descriptor algorithms to govern how an input region of an image is locally quantified. They take into account the locality of regions in an image and describe them separately. A HOG (histogram of gradients), as reported in [60,62,63], is a basic technique to extract a local description of the gradient magnitude and orientation of images. A HOG is invariant to photometric transformations but can only be used for human detection at fixed size. Scale-invariant feature transform (SIFT) was used in [64,65] and showed invariance to geometric and photometric transformation, even with 3D projection, but it contains high-dimensional features, which is computer inefficient and unsuitable for real-time applications. The speed-up robust feature (SURF) algorithm [66,67] is an alternative to SIFT that is faster and retains the detection points' quality. Lastly, the shape-based local feature descriptor in [68,69] demonstrated its robustness to noise by preserving the edge structures of the target subjects. However, it is heavily dependent on silhouette segmentation in its preprocessing step.
- Global representative features use an image descriptor that governs how an input image is globally quantified and returns a feature vector abstractly representing the image contents. The global descriptor in [31,70] encoded the detail information of corners, edges, ridges, and the optical flow as essential features. These features can be easily obtained from the camera depth but they are scene-dependent and have a lack of generic info. Some researchers in [71–73] used 3D space–time volume to extract 3D global feature vectors, which were independent of background subtraction. However, these 3D features were highly sensitive to noise and occlusion. The discrete Fourier transform (DFT) was also used in [74,75] to transform spatial features into frequency features, but the inverse process may lose the spatial and temporal information needed to identify the anomalous target subjects.
- The semantic features are obtained from the analysis of human body postures in a video sequence. These features can be transformed into human pose information [76,77] that is robust to interclass variation, but pose accuracy is hard to achieve. The researchers

in [78,79] used appearance-based features such as textures and colors to gain contextual information, but this was sensitive to intraclass variations. Three-dimensional semantic features can be obtained from RGB-D cameras [68,80], which provide both geometric and visual information, but this is highly affected by noise and occlusion problems.

5. Deep Learning Methods

This section discusses the deep learning methods used in AbHAR in video surveillance. Deep learning methods have achieved significant results, in term of both behavior recognition and video comprehension. A deep learning system can automatically learn and extract representative features from image sequences that contain both spatial and temporal information, using a AbHAR process. Based on different strategies, processing techniques, and the final objective of the network architecture, these deep learning methods can be categorized into four main groups; namely, reconstruction-based methods, multiclass classification methods, future frame prediction methods, and scoring methods. Table 1 summarizes the deep learning-based techniques used for anomaly detection in different categories.

Table 1. Summary of deep learning-based techniques for anomaly detection.

Method	Type	Year	Reference
Convolutional Long Short-Term Memory	RB + Future Frame Prediction	2016	[4]
2D Convolutional Autoencoder	RB	2016	[81]
Sparse Autoencoder	Reconstruction based	2016	[82]
Slow Feature Analysis + Deep Neural Network	Scoring	2016	[9]
Sparse Denoise Autoencoder	Multiclass Classification	2017	[83]
Autoencoder + Cascade Deep CNN	Multiclass Classification	2017	[3]
Spatiotemporal Autoencoder	RB + Future Frame Prediction	2017	[84]
Pretrained DNN + Gaussian classifier	Multiclass Classification	2018	[85]
Autoencoder + Low level features	Reconstruction based	2018	[86]
Multiple Instance Learning	Scoring	2018	[2]
Low-level Features + Autoencoder	Reconstruction based	2018	[86]
Frame predict using GANs	Future Frame Prediction	2018	[43]
Combination of traditional and deep features	Scoring	2019	[87]
Localization feature extraction	Scoring	2019	[88]
AnomalyNet	Reconstruction based	2019	[89]
Optical Flow + Multiple Instance Learning	Scoring	2019	[90]
Social Force Maps + Multiple Instance Learning	Scoring	2019	[91]
Attention module + Autoencoder	Reconstruction based	2019	[92]
Component Analysis + Transfer Learning	Multiclass Classification	2019	[93]
Object detection using SSD + Autoencoder	Multiclass Classification	2019	[94]
Sparse coding Deep neural network	Scoring	2019	[95]
Adaptive Intra-Frame Classification Network	Classification	2019	[96]
Autoencoder + Gaussian Mixture Model	Scoring	2020	[97]

5.1. Reconstruction-Based Method

This is the most common method used in AbHAR [82,92]. Formally, let x be the input image sequences and f be the neural network that reconstructs x . The reconstruction cost function Θ can be defined as a function to compute the error e between the original input x and $f(x)$, see Equation (1).

$$e = \Theta(x, f(x)) \quad (1)$$

A popular neural network used for the reconstruction error function Θ is an autoencoder network. An autoencoder network has two parts: an encoder and decoder; see Figure 12 for its schematic structure. The encoder is a neural network that has the capability to encode the input X into latent features that contain compact and discriminative representatives z . The decoder is also a neural network that has the ability to decode these representative features into their original form X' . The network is trained to reduce the error value e computed by the cost function Θ between X and X' .

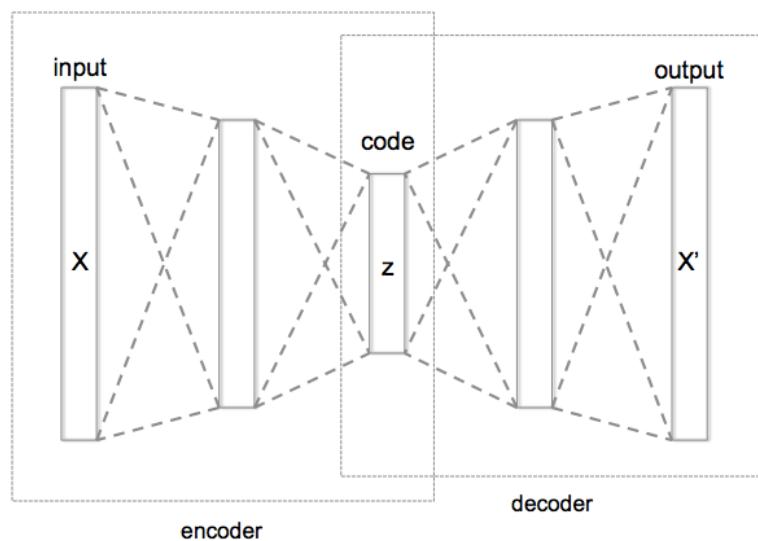


Figure 12. Autoencoder schematic structure [98].

Many approaches have utilized an autoencoder structure for identifying anomalies. The work in [81] used a sparse coding autoencoder to preserve the spatiotemporal information between the input and output. They used a 2D convolution network to encode the grayscale 2D image sequences from each segment of the videos. The authors only used one-channel input image stacking in the temporal dimension, for better reconstruction in both the spatial and temporal domains. The work of [4] also used a reconstruction-based method as their approach to AbHAR. However, they enhanced the spatiotemporal information using a convolutional long-short-term memory (Conv-LSTM) network as their main architecture. An LSTM network [99] has the ability to capture the long-term information of video data and hence increase the prediction accuracy. The researchers in [86] used low-level features such as edges and optical flow to add extra information to raw frames. Another use of autoencoders was proposed in [82], where the authors used two different autoencoders for their task: the first was a normal autoencoder, and the second was a sparse autoencoder that constrained the dimensional features and retained the most useful active neurons in the latent layer (see Figure 13).

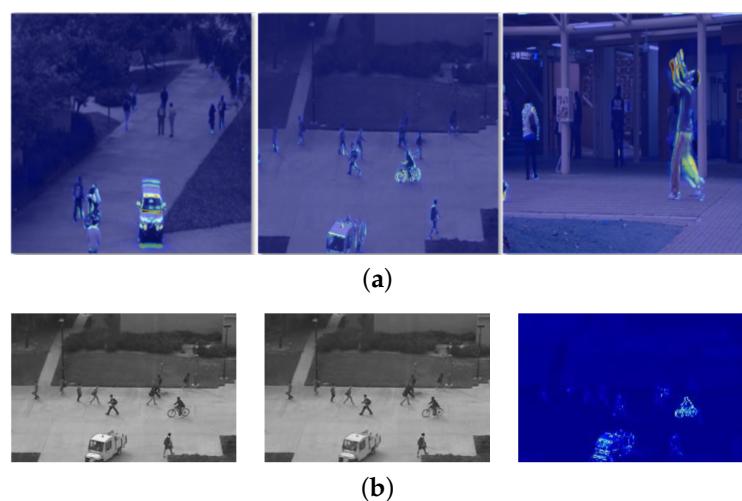


Figure 13. *Cont.*

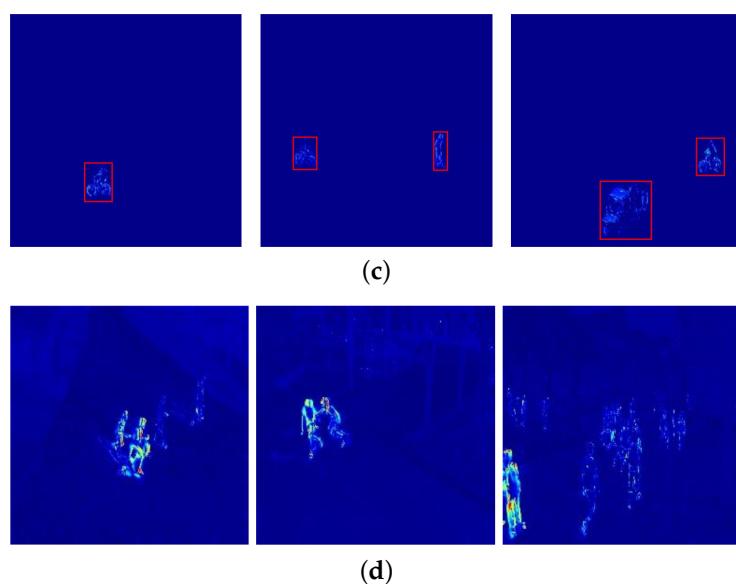


Figure 13. Error maps from recent methods. (a) Zero-shot cross-domain video anomaly detection (zxVAD)—Aich et al. [100] (2023); (b) dynamic local aggregation network—Yang et al. [101] (2022); (c) spatialtemporal memories—Liu et al. [102] (2022); (d) hybrid attention and motion constraint—Zhang et al. [103] (2022).

In [84], the authors used a 3D convolution autoencoder to preserve the temporal information and keep track of the spatial features in the temporal dimension. They also applied data augmentation, to increase the number of training samples. Later, the authors in [89] proposed a novel network for a feature learning framework that combined both motion and appearance features in an image, namely SC2NET. This network had the capability to compute sparsity loss and was trained through construction error, to learn how to construct useful spatiotemporal features.

The approach using an autoencoder is based on the assumption that a network will return a high construction error score for an abnormal instance. However, this assumption does not necessarily hold true, since, in some cases, autoencoders can generalize the abnormal instance as well as the normal instance. This means the reconstruction error score is lower than expected. The researchers in [92] proposed new approach to deal with this problem. They treated each encoding feature as a query for the decoder network. All normal encoding features were stored into memory. The decoder returned the closest normal encoding in the memory to each query instance. An abnormal instance hence cannot map the closest normal encoding and return a high reconstruction error (see Table 2).

Table 2. Comparison of recent methods used for video anomaly detection.

Method	Ped 1	Ped 2	Avenue	ShanghaiTech	Year	Reference
Cognitive memory-augmented network (CMAN)	-	96.2	-	-	2021	[104]
Single and multi-frame anomaly detection	-	97.5	87.2	-	2021	[105]
Multi-Level Memory modules in an Autoencoder with Skip Connections (ML-MemAE-SC)	-	99.3	91.1	76.2	2021	[106]
Autoencoder with a Memory Module (AMM) Explanation for Anomaly Detection	-	97.2	87.9	70.2	2021	[107]
Attention-based adversarial autoencoder (A3N) Group Activities for AD	73.1 90.7	80.1 97.7	- 89.4	- 86.9	2021 2022	[108] [109]
Variational Anomaly Detection Network (VADNet)	84.4	95.0	82.3	-	2022	[110]
	-	96.8	87.3	75.2	2022	[111]

Table 2. Cont.

Method	Ped 1	Ped 2	Avenue	ShanghaiTech	Year	Reference
Context-related video anomaly detection	-	96.3	87.1	73.6	2022	[112]
Localization based Reconstruction (LBR)	81.1	97.2	92.8	72.6	2022	[113]
Foreground–Background Separation Mutual						
Generative Adversarial Network (FSM-GAN)	-	98.1	80.1	73.5	2022	[114]
Dual-stream memory network	-	98.3	88.6	75.7	2023	[115]
Attention-based residual autoencoder	-	97.4	86.7	73.6	2023	[116]
Bi-directional Frame Interpolation	-	98.9	89.7	75.0	2023	[117]
Zero-shot Cross-domain Video Anomaly Detection (zxVAD)	78.6	95.8	83.2	-	2023	[100]

5.2. Multiclass Classifier Method

Despite the popularity of using error reconstruction as the main method for anomaly detection. Different approaches have framed the anomaly detection problem as a multiclass classification. This classification method receives the video segment as an input x and returns an output y indicating a class label in redefined categories, see Equation (2).

$$y = f(x), y \in \mathbb{R} \quad (2)$$

One of the main problems in anomaly detection is imbalanced datasets, and the approach in [3] tried to resolve this problem using a cascade framework to learn compact and robust features. There are two stages in this approach, the first uses a stack of autoencoders to only learn from the normal video patches. The second stage is a convolutional neural network, which processes the video patches that cannot be modeled in the first stage and need further investigation. The final features are put into a Gaussian classifier to perform classification. Another approach seen in [83] used both local and global descriptors to learn the representative features from a video segment. These features were composed of both spatial and temporal information from each video patch. For a local descriptor, they used a similarity metric to extract spatiotemporal features. Meanwhile, the global descriptor utilized a pretrained autoencoder network to learn latent features. Next, these features were put into another autoencoder to select the most representative ones and the selected features were fed into a Gaussian classifier to detect anomalies. The authors in [85] also used Gaussian classifiers to make a prediction; however, they took advantage of pretrained models and extracted intermediate features, to find robust patterns for the classifiers to detect anomalies. In cases where the classifier could not make a decision, the features were sent to the top layers in the pretrained models, to learn more discriminative features.

The work of [93] also used pretrained models to learn representative features. They explored the feature spaces of a convolutional neural network in different domains, using the technique of transfer component analysis [118] to make a generalization. The method proposed in [94] focused on each human object in a scene using a single-shot detector model (SSD) [119] for each frame in the video. After locating each target subject in the scene, two different autoencoder models were used to learn both motion and appearance features for each subject. These features were concatenated and clustered into different subsets using a k-mean technique [120]. Each cluster represented one kind of normal event. A one-versus-rest classifier was trained to perform classification of each cluster. Any event that received a negative score from the classifier and did not belong to any cluster was tagged as an abnormal event. Ref. [96] also framed the anomaly detection problem as a multiclass classification problem, by combining all the representative features from the optical flow and gradient in each video patch. This allowed the author to build an adaptive intraframe classification network that learned both motion and appearance features, to detect and localize anomalies in video patches.

5.3. Future Frame Prediction Method

One problem with reconstruction-based methods is that the autoencoder can accidentally reconstruct abnormal instances as well as normal ones. This may produce many false negatives and reduce the performance. The work of [29] suggested a new method, named future frame prediction, that can handle this problem well. Formally, given x_t is the input video segment at time step t , the future frame prediction method provides a function p to predict the next segment frames at time $t + 1$ and compare the error cost between the predicted frames and the current frames at that time. If the error value is greater than a defined threshold value, then the frames are tagged as abnormal instances. Equation (3) gives the mathematical form for this approach.

$$x_{t+1} = p(x_t) \quad (3)$$

A neural network called generative adversarial networks (GANs) [43] is used for this approach. Basically, GANs is a generative model that comprises two main parts: a generator, and a discriminator. The generator is used for generating new instance data based on statistical information from the training data. The discriminator's job is to verify whether the input is coming the generator (fake) or coming from training data (real). Ref. [121] used GANs in their approach. The generator was a neural network using a U-Net architecture [122] to predict future frames. The authors chose the U-Net model since it has shown significant performance in image-to-image translation tasks. The discriminator was a neural network used to determine the frames having abnormal events or not.

Some previous works on reconstruction-based methods also took advantage of this future frame prediction approach. For example, the approach in [81] encoded both spatial and temporal information from video segments and further predicted the future frame given a center frame. Some earlier works, such as ref. [84] and ref. [4], also leveraged the idea of future frame prediction to reconstruct the current frame. Reference [84] used two separated branches to learn how to predict the future frames and reconstruct the frame at the same time. The author used the loss value to train the network to extract temporal information. The loss value was a combination of both the prediction loss and the reconstruction loss. The approach in [4] also used a parallel branch to learn both frame prediction and reconstruction. However, their network could identify the point of interest within a video, to extract the useful information for detecting anomalies.

5.4. Scoring Method

In the scoring method, the network tries to predict the anomaly score for each video segment. Thus, it can be considered a regression problem, where the purpose is to assign a high score value (or sometimes a low score value) for any abnormal instance. Mathematically, the scoring method uses a function s to take the input video segments x and assign a t value indicating the anomaly score in each segment, see Equation (4).

$$t = s(x), t \in \mathbb{R} \quad (4)$$

The approach in [9] used a sum-squared derivative to calculate the anomaly score from the extracted features. These features comprised both deep and slow feature analysis [123], which was used to learn the semantic meaning from image sequences. The work in [2] used the technique of multiple-instance learning to measure the anomaly score from weakly labeled video segments. They utilized a 3D convolutional neural network [124] to learn both spatial and temporal information from the image sequences. The spatiotemporal features were fed into the network model. The model was trained with a supervised method, with weak labels to detect the final score for each video segment. The idea of using multiple-instance learning inspired the work in [90], where they used an optical flow feature along with an attention mechanism [125] to determine the promising features in video segments, to calculate the final score. Similarly based on an idea in [2], reference [91] proposed a new dual network to learn motion features by modifying the original model

structure. The authors added an attention module and utilized social fore maps [39] to learn motion representatives.

The authors in [87] used optical flow to foreground objects from the background video. Then a convolutional neural network was used to extract the identified objects' features, along with a histogram of optical flow. All the features were combined using a matrix factorization method and clustered into different subsets. A voting system detected whether new instances were normal or abnormal. The approach in [88] focused on localization in video segments, by extracting a tube from the video. They also showed that their method could accurately identify anomalies based on scoring values.

The researchers in [95] used sparse coding for anomaly detection. The sparse coding [41] first learned the dictionary mapping from all normal instances. The authors in [95] proposed temporal-coherent sparse coding using both spatial and temporal information extracted from a stacked recurrent neural network autoencoder. The work in [97] used a Gaussian mixture model for anomaly detection. They proposed a new network, named Gaussian mixture fully convolutional variational autoencoder, to model video patches using both motion and appearance features. The model predicted the probability score for each video patch through a simple energy-based method.

5.5. Anomaly Score

The difference between the predicted frame/reconstructed frame \hat{I} and the ground truth frame I is computed using the peak signal-to-noise ratio (PSNR):

$$\text{PSNR}(I, \hat{I}) = 10 \log_{10} \frac{[\max_{\hat{I}}]^2}{\frac{1}{N} \sum_{i=1}^N (I_i - \hat{I}_i)^2} \quad (5)$$

The PSNR scores of all frames in a video are normalized to the range $[0, 1]$. Then, the anomaly score is obtained using the following formula:

$$S(t) = \frac{\text{PSNR}_t - \min(\text{PSNR})}{\max(\text{PSNR}) - \min(\text{PSNR})} \quad (6)$$

The anomaly score $S(t)$ is used to determine whether a frame is a normal or abnormal event.

6. Research Gaps, Challenges, and Future Research

6.1. Research Gaps

Recent methods [43,126,127] have focused on exploiting both the appearance and motion information in video, by extracting structural features and the optical flow. The extraction of optical flow requires a high computational cost. Therefore, some proposed methods [116,128,129] have tried to capture motion information without using optical flow.

On the other hand, most recently proposed methods have tried to improve the accuracy of anomaly detection systems by applying a modern model [130,131]. We need to consider the complexity of the model, since an anomaly detection system should be run as a real-time application.

6.2. Challenges

This section discusses the general challenges in AbHAR for video surveillance, as well as possible solutions and future research. Overall, researchers have relied on supervised learning methods for their approach. This requires a large amount of labeled data for training and improving the model accuracy. However, collecting data is a tedious and laborious process. There is also no standard dataset that can cover all abnormal behaviors in real situations. These problems can be solved by utilizing transfer learning techniques or crowd-sourcing [132]. In addition, there is a need for standard metrics that allow a fair comparison between different approaches [133].

During inference, there are also some limitations in most approaches, such as false alarms due to subtle details in human motion and appearance. In some cases, early

detection is also necessary for crime prevention. These challenges can be approached by choosing representative features for AbHAR [134] that allow the surveillance system to learn and simulate human behaviors. Intra-class variation and inter-class similarity problems occur when a model considers non-similar patterns between the same activity or similar representatives among different classes. This can lead to an ambiguous boundary between the normal and abnormal events, and requires the model to learn only unique and discriminative features for classification [135]. Moreover, the tracking of multi-subject interactions in a group of people is a challenging problem that requires a model to have the ability to capture spatiotemporal information from subjects [136].

A noisy environment is considered the main issue in many surveillance systems. A cluttered and dynamic background is challenging to model accurately. In addition, occlusion, low illumination, low-quality videos, and various viewpoints also occur in real-time surveillance. Multimodal data can overcome these challenges [137,138], e.g., the usage of D-RGB images to extract depth information from a video segment. Moreover, a surveillance system requires real-time sensing, which is highly energy consuming and requires a lot of computing power. This raises the requirement of using adaptive methods based on the principle of a sampling frequency and iterative segmentation [139].

6.3. Future Research

Future research will focus on using transfer learning methods in deep learning approaches, [140] such as pretrained models to explore the spatiotemporal relationships in video segments. In the context of AbHAR, transfer learning can be used to transfer the knowledge gained from a large-scale dataset to a small-scale dataset, where the latter may have different distributions and challenges. Future research could investigate the effectiveness of transfer learning in AbHAR using deep learning-based methods. Researchers currently pay more attention of how to interpret a video model by selecting key frame information [64,77] to classify activity classes. Moreover, extracting rich semantic features from multimodal datasets is a research trend [75,141], since this information enables explaining the long-term relationship between interacting objects [142]. Lastly, the physical interactions between humans or humans–objects [143] require more investigation, since these interactions can provide useful features for understanding human behaviors and provide more interpretive features for model training.

7. Conclusions

Video surveillance systems are becoming increasingly important in detecting and preventing criminal activities in various settings, such as public spaces, offices, and homes. One of the essential components of a video surveillance system is abnormal human activity recognition (AbHAR). AbHAR is critical for detecting and understanding anomalous behaviors, which could be potential threats to the safety and security of the environment being monitored. This paper provides a comprehensive review of deep learning-based methods in video surveillance. Human activity recognition plays an important role in detecting and understanding anomalous behaviors. This review has covered the essential aspects of related works and background knowledge. It has also presented different benchmark databases used in AbHAR, with a detailed description of each one. These databases are increasing in both quantity and quality, allowing researchers to study many aspects related to real-time scenarios. This comprehensive review of deep learning-based methods has introduced methods such as reconstruction based approaches, multiclass classification, future frame prediction, and scoring methods. This diversity of approaches demonstrates the difficulty of AbHAR, which has forced researchers to find better solutions. In addition, future studies on the transfer learning method or multimodal datasets will provide initial steps toward a better analysis in anomaly detection. In fact, the choice of representative features in the feature engineering step is the most difficult part in any surveillance system. Robust feature extraction from a dynamic background is an essential

part of capturing relative and discriminative patterns for tracking and understanding human behaviors.

Author Contributions: All authors contributed to the study conceptualization and design. The first draft of the manuscript was written by H.-T.D. and V.-T.L., supervised by V.T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Ho Chi Minh City Open University, Vietnam. Portions of the research in this paper used the CASIA Action Database collected by the Institute of Automation, Chinese Academy of Sciences.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

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

References

- Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 1–58. [[CrossRef](#)]
- 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.
- Sabokrou, M.; Fayyaz, M.; Fathy, M.; Klette, R. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Trans. Image Process.* **2017**, *26*, 1992–2004. [[CrossRef](#)] [[PubMed](#)]
- Medel, J.R.; Savakis, A. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv* **2016**, arXiv:1612.00390.
- Hu, C.; Chen, Y.; Hu, L.; Peng, X. A novel random forests based class incremental learning method for activity recognition. *Pattern Recognit.* **2018**, *78*, 277–290. [[CrossRef](#)]
- Xiao, Q.; Song, R. Action recognition based on hierarchical dynamic Bayesian network. *Multimed. Tools Appl.* **2018**, *77*, 6955–6968. [[CrossRef](#)]
- Sok, P.; Xiao, T.; Azeze, Y.; Jayaraman, A.; Albert, M.V. Activity recognition for incomplete spinal cord injury subjects using hidden Markov models. *IEEE Sens. J.* **2018**, *18*, 6369–6374. [[CrossRef](#)]
- Abidine, B.M.; Fergani, L.; Fergani, B.; Oussalah, M. The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition. *Pattern Anal. Appl.* **2018**, *21*, 119–138. [[CrossRef](#)]
- Hu, X.; Hu, S.; Huang, Y.; Zhang, H.; Wu, H. Video anomaly detection using deep incremental slow feature analysis network. *IET Comput. Vis.* **2016**, *10*, 258–267. [[CrossRef](#)]
- Ijjina, E.P.; Chalavadi, K.M. Human action recognition in RGB-D videos using motion sequence information and deep learning. *Pattern Recognit.* **2017**, *72*, 504–516. [[CrossRef](#)]
- Tan, T.H.; Gochoo, M.; Huang, S.C.; Liu, Y.H.; Liu, S.H.; Huang, Y.F. Multi-resident activity recognition in a smart home using RGB activity image and DCNN. *IEEE Sens. J.* **2018**, *18*, 9718–9727. [[CrossRef](#)]
- Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [[CrossRef](#)]
- Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
- Oyedotun, O.K.; Khashman, A. Deep learning in vision-based static hand gesture recognition. *Neural Comput. Appl.* **2017**, *28*, 3941–3951. [[CrossRef](#)]
- Weingaertner, T.; Hassfeld, S.; Dillmann, R. Human motion analysis: A review. In Proceedings of the IEEE Nonrigid and Articulated Motion Workshop, San Juan, PR, USA, 16 June 1997; p. 0090.
- Aggarwal, J.K.; Cai, Q. Human motion analysis: A review. *Comput. Vis. Image Underst.* **1999**, *73*, 428–440. [[CrossRef](#)]
- Valera, M.; Velastin, S.A. Intelligent distributed surveillance systems: A review. *IEEE Proc. Vision Image Signal Process.* **2005**, *152*, 192–204. [[CrossRef](#)]
- Ko, T. A survey on behavior analysis in video surveillance for homeland security applications. In Proceedings of the 2008 37th IEEE Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, 15–17 October 2008; pp. 1–8.
- Popoola, O.P.; Wang, K. Video-based abnormal human behavior recognition—A review. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2012**, *42*, 865–878. [[CrossRef](#)]
- Tsitsoulis, A.; Bourbakis, N. A first stage comparative survey on vision-based human activity recognition. *Int. J. Tools* **2013**, *24*, 782783783.
- Zablocki, M.; Gościewska, K.; Frejlichowski, D.; Hofman, R. Intelligent video surveillance systems for public spaces—A survey. *J. Theor. Appl. Comput. Sci.* **2014**, *8*, 13–27.
- Li, T.; Chang, H.; Wang, M.; Ni, B.; Hong, R.; Yan, S. Crowded scene analysis: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *25*, 367–386. [[CrossRef](#)]
- Onofri, L.; Soda, P.; Pechenizkiy, M.; Iannello, G. A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Syst. Appl.* **2016**, *63*, 97–111. [[CrossRef](#)]

24. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
25. Mabrouk, A.B.; Zagrouba, E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Syst. Appl.* **2018**, *91*, 480–491. [CrossRef]
26. Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; Escalera, S. RGB-D-based human motion recognition with deep learning: A survey. *Comput. Vis. Image Underst.* **2018**, *171*, 118–139. [CrossRef]
27. Abdallah, Z.S.; Gaber, M.M.; Srinivasan, B.; Krishnaswamy, S. Activity recognition with evolving data streams: A review. *ACM Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
28. 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]
29. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. *Sensors* **2019**, *19*, 1005. [CrossRef]
30. Raval, R.M.; Prajapati, H.B.; Dabhi, V.K. Survey and analysis of human activity recognition in surveillance videos. *Intell. Decis. Technol.* **2019**, *13*, 271–294. [CrossRef]
31. Dhiman, C.; Vishwakarma, D.K. A review of state-of-the-art techniques for abnormal human activity recognition. *Eng. Appl. Artif. Intell.* **2019**, *77*, 21–45. [CrossRef]
32. Ramachandra, B.; Jones, M.J.; Vatsavai, R.R. A Survey of Single-Scene Video Anomaly Detection. *arXiv* **2020**, arXiv:2004.05993.
33. Dang, L.M.; Min, K.; Wang, H.; Piran, M.J.; Lee, C.H.; Moon, H. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognit.* **2020**, *108*, 107561. [CrossRef]
34. Ahmed, M.; Mahmood, A.N.; Hu, J. A survey of network anomaly detection techniques. *J. Netw. Comput. Appl.* **2016**, *60*, 19–31. [CrossRef]
35. Moustafa, N.; Hu, J.; Slay, J. A holistic review of network anomaly detection systems: A comprehensive survey. *J. Netw. Comput. Appl.* **2019**, *128*, 33–55. [CrossRef]
36. Fernandes, G.; Rodrigues, J.J.; Carvalho, L.F.; Al-Muhtadi, J.F.; Proen  a, M.L. A comprehensive survey on network anomaly detection. *Telecommun. Syst.* **2019**, *70*, 447–489. [CrossRef]
37. CASIA. CASIA Action Database. 2007. Available online: <http://www.cbsr.ia.ac.cn/english/Action/%20Databases/%20EN.asp> (accessed on 20 November 2020).
38. 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]
39. 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.
40. Zaharescu, A.; Wildes, R. Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In Proceedings of the European Conference on Computer Vision, Heraklion Crete, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 563–576.
41. Lu, C.; Shi, J.; Jia, J. Abnormal event detection at 150 FPS in MATLAB. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2720–2727.
42. Statistical Visual Computing Lab. UCSD Anomaly Data Set. 2014. Available online: <http://www.svcl.ucsd.edu/projects/anomaly/> (accessed on 20 November 2020).
43. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future frame prediction for anomaly detection—A new baseline. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6536–6545.
44. Ramachandra, B.; Jones, M. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass, CO, USA, 1–5 March 2020; pp. 2569–2578.
45. Babaee, M.; Dinh, D.T.; Rigoll, G. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognit.* **2018**, *76*, 635–649. [CrossRef]
46. Subudhi, B.N.; Ghosh, S.; Shiu, S.C.; Ghosh, A. Statistical feature bag based background subtraction for local change detection. *Inf. Sci.* **2016**, *366*, 31–47. [CrossRef]
47. Shen, Y.; Hu, W.; Yang, M.; Liu, J.; Wei, B.; Lucey, S.; Chou, C.T. Real-time and robust compressive background subtraction for embedded camera networks. *IEEE Trans. Mob. Comput.* **2015**, *15*, 406–418. [CrossRef]
48. Jiang, S.; Lu, X. WeSamBE: A weight-sample-based method for background subtraction. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2105–2115. [CrossRef]
49. Sakkos, D.; Liu, H.; Han, J.; Shao, L. End-to-end video background subtraction with 3d convolutional neural networks. *Multimed. Tools Appl.* **2018**, *77*, 23023–23041. [CrossRef]
50. Minematsu, T.; Shimada, A.; Uchiyama, H.; Taniguchi, R.I. Analytics of deep neural network-based background subtraction. *J. Imaging* **2018**, *4*, 78. [CrossRef]
51. Ullah, A.; Muhammad, K.; Del Ser, J.; Baik, S.W.; de Albuquerque, V.H.C. Activity recognition using temporal optical flow convolutional features and multilayer LSTM. *IEEE Trans. Ind. Electron.* **2018**, *66*, 9692–9702. [CrossRef]
52. Ladjailia, A.; Bouchrika, I.; Merouani, H.F.; Harrati, N.; Mahfouf, Z. Human activity recognition via optical flow: Decomposing activities into basic actions. *Neural Comput. Appl.* **2020**, *32*, 16387–16400. [CrossRef]
53. Mliki, H.; Bouhlel, F.; Hammami, M. Human activity recognition from UAV-captured video sequences. *Pattern Recognit.* **2020**, *100*, 107140. [CrossRef]

54. Singh, R.; Dhillon, J.K.; Kushwaha, A.K.S.; Srivastava, R. Depth based enlarged temporal dimension of 3D deep convolutional network for activity recognition. *Multimed. Tools Appl.* **2019**, *78*, 30599–30614. [[CrossRef](#)]
55. Liu, L.; Wang, S.; Peng, Y.; Huang, Z.; Liu, M.; Hu, B. Mining intricate temporal rules for recognizing complex activities of daily living under uncertainty. *Pattern Recognit.* **2016**, *60*, 1015–1028. [[CrossRef](#)]
56. Sun, B.; Li, Y.; Guosheng, C.; Zhang, J.; Chang, B.; Min, C. Moving target segmentation using Markov random field-based evaluation metric in infrared videos. *Opt. Eng.* **2018**, *57*, 013106. [[CrossRef](#)]
57. Lai, Y.; Han, Y.; Wang, Y. Anomaly detection with prototype-guided discriminative latent embeddings. In Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM), Auckland, New Zealand, 7–10 December 2021; pp. 300–309.
58. Doshi, K.; Yilmaz, Y. Any-shot sequential anomaly detection in surveillance videos. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 934–935.
59. 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, Online, 2–9 February 2021; Volume 35, pp. 938–946.
60. Kumar, K.S.; Bhavani, R. Human activity recognition in egocentric video using HOG, GiST and color features. *Multimed. Tools Appl.* **2020**, *79*, 3543–3559. [[CrossRef](#)]
61. Wang, Z.; Wu, D.; Chen, J.; Ghoneim, A.; Hossain, M.A. A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection. *IEEE Sens. J.* **2016**, *16*, 3198–3207. [[CrossRef](#)]
62. Roy, P.K.; Om, H. Suspicious and violent activity detection of humans using HOG features and SVM classifier in surveillance videos. In *Advances in Soft Computing and Machine Learning in Image Processing*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 277–294.
63. Patel, C.I.; Garg, S.; Zaveri, T.; Banerjee, A.; Patel, R. Human action recognition using fusion of features for unconstrained video sequences. *Comput. Electr. Eng.* **2018**, *70*, 284–301. [[CrossRef](#)]
64. Jagadeesh, B.; Patil, C.M. Video based human activity detection, recognition and classification of actions using SVM. *Trans. Mach. Learn. Artif. Intell.* **2018**, *6*, 22–22.
65. Kale, G.V. Human Activity Recognition on Real Time and Offline Dataset. *Int. J. Intell. Syst. Appl. Eng.* **2019**, *7*, 60–65. [[CrossRef](#)]
66. Bux, A.; Angelov, P.; Habib, Z. Vision based human activity recognition: A review. In *Advances in Computational Intelligence Systems*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 341–371.
67. Ahad, M.A.R.; Tan, J.K.; Kim, H.; Ishikawa, S. Activity representation by SURF-based templates. *Comput. Methods Biomed. Biomed. Eng. Imaging Vis.* **2018**, *6*, 573–583. [[CrossRef](#)]
68. Patrona, F.; Chatzifotis, A.; Zarpalas, D.; Daras, P. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognit.* **2018**, *76*, 612–622. [[CrossRef](#)]
69. Tang, J.; Cheng, H.; Zhao, Y.; Guo, H. Structured dynamic time warping for continuous hand trajectory gesture recognition. *Pattern Recognit.* **2018**, *80*, 21–31. [[CrossRef](#)]
70. Kim, K.; Jalal, A.; Mahmood, M. Vision-based Human Activity recognition system using depth silhouettes: A Smart home system for monitoring the residents. *J. Electr. Eng. Technol.* **2019**, *14*, 2567–2573. [[CrossRef](#)]
71. Baumann, F.; Ehlers, A.; Rosenhahn, B.; Liao, J. Recognizing human actions using novel space-time volume binary patterns. *Neurocomputing* **2016**, *173*, 54–63. [[CrossRef](#)]
72. Kihl, O.; Picard, D.; Gosselin, P.H. Local polynomial space–time descriptors for action classification. *Mach. Vis. Appl.* **2016**, *27*, 351–361. [[CrossRef](#)]
73. Fu, Y.; Zhang, T.; Wang, W. Sparse coding-based space-time video representation for action recognition. *Multimed. Tools Appl.* **2017**, *76*, 12645–12658. [[CrossRef](#)]
74. Seyfioğlu, M.S.; Özbayoglu, A.M.; Gürbüz, S.Z. Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities. *IEEE Trans. Aerosp. Electron. Syst.* **2018**, *54*, 1709–1723. [[CrossRef](#)]
75. Shahroudy, A.; Ng, T.T.; Yang, Q.; Wang, G. Multimodal multipart learning for action recognition in depth videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2123–2129. [[CrossRef](#)]
76. Li, M.; Zhou, Z.; Liu, X. Multi-person pose estimation using bounding box constraint and LSTM. *IEEE Trans. Multimed.* **2019**, *21*, 2653–2663. [[CrossRef](#)]
77. Nishi, K.; Miura, J. Generation of human depth images with body part labels for complex human pose recognition. *Pattern Recognit.* **2017**, *71*, 402–413. [[CrossRef](#)]
78. 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](#)]
79. Zhang, Y.; Lu, H.; Zhang, L.; Ruan, X. Combining motion and appearance cues for anomaly detection. *Pattern Recognit.* **2016**, *51*, 443–452. [[CrossRef](#)]
80. Ji, X.; Cheng, J.; Feng, W.; Tao, D. Skeleton embedded motion body partition for human action recognition using depth sequences. *Signal Process.* **2018**, *143*, 56–68. [[CrossRef](#)]
81. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning temporal regularity in video sequences. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 733–742.
82. 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](#)]

83. Narasimhan, M.G.; Kamath, S. Dynamic video anomaly detection and localization using sparse denoising autoencoders. *Multimed. Tools Appl.* **2018**, *77*, 13173–13195. [CrossRef]
84. 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.
85. Sabokrou, M.; Fayyaz, M.; Fathy, M.; Moayed, Z.; Klette, R. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Comput. Vis. Image Underst.* **2018**, *172*, 88–97. [CrossRef]
86. Ribeiro, M.; Lazzaretti, A.E.; Lopes, H.S. A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognit. Lett.* **2018**, *105*, 13–22. [CrossRef]
87. Sabzalian, B.; Marvi, H.; Ahmadyfard, A. Deep and Sparse features For Anomaly Detection and Localization in video. In Proceedings of the 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), Tehran, Iran, 6–7 March 2019; pp. 173–178.
88. Landi, F.; Snoek, C.G.; Cucchiara, R. Anomaly Locality in Video Surveillance. *arXiv* **2019**, arXiv:1901.10364.
89. Zhou, J.T.; Du, J.; Zhu, H.; Peng, X.; Liu, Y.; Goh, R.S.M. AnomalyNet: An anomaly detection network for video surveillance. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2537–2550. [CrossRef]
90. Zhu, Y.; Newsam, S. Motion-aware feature for improved video anomaly detection. *arXiv* **2019**, arXiv:1907.10211.
91. Lin, S.; Yang, H.; Tang, X.; Shi, T.; Chen, L. Social MIL: Interaction-Aware for Crowd Anomaly Detection. In Proceedings of the 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 18–21 September 2019; pp. 1–8.
92. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; van den Hengel, A. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1705–1714.
93. dos Santos, F.P.; Ribeiro, L.S.; Ponti, M.A. Generalization of feature embeddings transferred from different video anomaly detection domains. *J. Vis. Commun. Image Represent.* **2019**, *60*, 407–416. [CrossRef]
94. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7842–7851.
95. Luo, W.; Liu, W.; Lian, D.; Tang, J.; Duan, L.; Peng, X.; Gao, S. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1070–1084. [CrossRef] [PubMed]
96. Xu, K.; Sun, T.; Jiang, X. Video anomaly detection and localization based on an adaptive intra-frame classification network. *IEEE Trans. Multimed.* **2019**, *22*, 394–406. [CrossRef]
97. Fan, Y.; Wen, G.; Li, D.; Qiu, S.; Levine, M.D.; Xiao, F. Video anomaly detection and localization via Gaussian mixture fully convolutional variational autoencoder. *Comput. Vis. Image Underst.* **2020**, *195*, 102920. [CrossRef]
98. Wikipedia. Autoencoder—Wikipedia, The Free Encyclopedia. 2023. Available online: <http://en.wikipedia.org/w/index.php?title=Autoencoder&oldid=1141727025> (accessed on 4 April 2023).
99. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
100. Aich, A.; Peng, K.C.; Roy-Chowdhury, A.K. Cross-Domain Video Anomaly Detection without Target Domain Adaptation. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2579–2591.
101. Yang, Z.; Wu, P.; Liu, J.; Liu, X. Dynamic Local Aggregation Network with Adaptive Clusterer for Anomaly Detection. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part IV; Springer: Berlin/Heidelberg, Germany, 2022; pp. 404–421.
102. Liu, Y.; Liu, J.; Zhao, M.; Yang, D.; Zhu, X.; Song, L. Learning Appearance-Motion Normality for Video Anomaly Detection. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; pp. 1–6.
103. Zhang, X.; Fang, J.; Yang, B.; Chen, S.; Li, B. Hybrid Attention and Motion Constraint for Anomaly Detection in Crowded Scenes. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 2259–2274. [CrossRef]
104. Wang, T.; Xu, X.; Shen, F.; Yang, Y. A cognitive memory-augmented network for visual anomaly detection. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 1296–1307. [CrossRef]
105. Bahrami, M.; Pourahmadi, M.; Vafaei, A.; Shayesteh, M.R. A comparative study between single and multi-frame anomaly detection and localization in recorded video streams. *J. Vis. Commun. Image Represent.* **2021**, *79*, 103232. [CrossRef]
106. Liu, Z.; Nie, Y.; Long, C.; Zhang, Q.; Li, G. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 13588–13597.
107. Tang, W.; Feng, Y.; Li, J. An autoencoder with a memory module for video anomaly detection. In Proceedings of the 2021 36th Youth Academic Annual Conference of Chinese Association of Automation (YAC), Nanchang, China, 28–30 May 2021; pp. 473–478.
108. Wu, C.; Shao, S.; Tunc, C.; Satam, P.; Hariri, S. An explainable and efficient deep learning framework for video anomaly detection. *Clust. Comput.* **2022**, *25*, 2715–2737. [CrossRef]
109. Aslam, N.; Rai, P.K.; Kolekar, M.H. A3N: Attention-based adversarial autoencoder network for detecting anomalies in video sequence. *J. Vis. Commun. Image Represent.* **2022**, *87*, 103598. [CrossRef]

110. Borja-Borja, L.F.; Azorín-López, J.; Saval-Calvo, M.; Fuster-Guilló, A.; Sebban, M. Architecture for Automatic Recognition of Group Activities Using Local Motions and Context. *IEEE Access* **2022**, *10*, 79874–79889. [[CrossRef](#)]
111. Huang, X.; Hu, Y.; Luo, X.; Han, J.; Zhang, B.; Cao, X. Boosting Variational Inference with Margin Learning for Few-Shot Scene-Adaptive Anomaly Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *early access*. [[CrossRef](#)]
112. Li, D.; Nie, X.; Li, X.; Zhang, Y.; Yin, Y. Context-related video anomaly detection via generative adversarial network. *Pattern Recognit. Lett.* **2022**, *156*, 183–189. [[CrossRef](#)]
113. Yu, G.; Wang, S.; Cai, Z.; Liu, X.; Xu, C.; Wu, C. Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13987–13998.
114. Zhang, Z.; Zhong, S.h.; Fares, A.; Liu, Y. Detecting abnormality with separated foreground and background: Mutual generative adversarial networks for video abnormal event detection. *Comput. Vis. Image Underst.* **2022**, *219*, 103416. [[CrossRef](#)]
115. Wang, Z.; Chen, Y. Anomaly detection with dual-stream memory network. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103739. [[CrossRef](#)]
116. Le, V.T.; Kim, Y.G. Attention-based residual autoencoder for video anomaly detection. *Appl. Intell.* **2023**, *53*, 3240–3254. [[CrossRef](#)]
117. Deng, H.; Zhang, Z.; Zou, S.; Li, X. Bi-Directional Frame Interpolation for Unsupervised Video Anomaly Detection. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 2634–2643.
118. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw.* **2010**, *22*, 199–210. [[CrossRef](#)]
119. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
120. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 21 June–18 July 1965; Volume 1, pp. 281–297.
121. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Cambridge, MA, USA, 8–13 December 2014; pp. 2672–2680.
122. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Cham, Switzerland, 2015; pp. 234–241.
123. Wiskott, L.; Sejnowski, T.J. Slow feature analysis: Unsupervised learning of invariances. *Neural Comput.* **2002**, *14*, 715–770. [[CrossRef](#)]
124. 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.
125. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
126. Wu, P.; Liu, J.; Li, M.; Sun, Y.; Shen, F. Fast sparse coding networks for anomaly detection in videos. *Pattern Recognit.* **2020**, *107*, 107515. [[CrossRef](#)]
127. Li, N.; Chang, F.; Liu, C. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE Trans. Multimed.* **2020**, *23*, 203–215. [[CrossRef](#)]
128. Chang, Y.; Tu, Z.; Xie, W.; Yuan, J. Clustering driven deep autoencoder for video anomaly detection. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 329–345.
129. Chang, Y.; Tu, Z.; Xie, W.; Luo, B.; Zhang, S.; Sui, H.; Yuan, J. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognit.* **2022**, *122*, 108213. [[CrossRef](#)]
130. Huang, X.; Zhao, C.; Gao, C.; Chen, L.; Wu, Z. Synthetic Pseudo Anomalies for Unsupervised Video Anomaly Detection: A Simple yet Efficient Framework based on Masked Autoencoder. *arXiv* **2023**, arXiv:2303.05112.
131. Sun, X.; Chen, J.; Shen, X.; Li, H. Transformer with Spatio-Temporal Representation for Video Anomaly Detection. In Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+ SSPR 2022, Montreal, QC, Canada, 26–27 August 2022; Springer: Berlin/Heidelberg, Germany, 2023; pp. 213–222.
132. Guo, L.; Wang, L.; Liu, J.; Zhou, W.; Lu, B. HuAc: Human activity recognition using crowdsourced WiFi signals and skeleton data. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 6163475. [[CrossRef](#)]
133. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970.
134. Pei, Z.; Qi, X.; Zhang, Y.; Ma, M.; Yang, Y.H. Human trajectory prediction in crowded scene using social-affinity Long Short-Term Memory. *Pattern Recognit.* **2019**, *93*, 273–282. [[CrossRef](#)]
135. Akila, K.; Chitrakala, S. Highly refined human action recognition model to handle intraclass variability & interclass similarity. *Multimed. Tools Appl.* **2019**, *78*, 20877–20894.

136. Wateosot, C.; Suvonvorn, N. Group activity recognition with an interaction force based on low-level features. *IEEE Trans. Electr. Electron. Eng.* **2019**, *14*, 1061–1073. [[CrossRef](#)]
137. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern Recognit.* **2017**, *61*, 295–308. [[CrossRef](#)]
138. Carreira, J.; Noland, E.; Hillier, C.; Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv* **2019**, arXiv:1907.06987.
139. Xu, C.; Chai, D.; He, J.; Zhang, X.; Duan, S. InnoHAR: A deep neural network for complex human activity recognition. *IEEE Access* **2019**, *7*, 9893–9902. [[CrossRef](#)]
140. Salaken, S.M.; Khosravi, A.; Nguyen, T.; Nahavandi, S. Seeded transfer learning for regression problems with deep learning. *Expert Syst. Appl.* **2019**, *115*, 565–577. [[CrossRef](#)]
141. Ordóñez, F.J.; Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* **2016**, *16*, 115. [[CrossRef](#)] [[PubMed](#)]
142. Nguyen, T.N.; Thai, C.H.; Luu, A.T.; Nguyen-Xuan, H.; Lee, J. NURBS-based postbuckling analysis of functionally graded carbon nanotube-reinforced composite shells. *Comput. Methods Appl. Mech. Eng.* **2019**, *347*, 983–1003. [[CrossRef](#)]
143. Goyal, R.; Kahou, S.E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The “Something Something” Video Database for Learning and Evaluating Visual Common Sense. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; Volume 1, p. 5.

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