applied sciences

MDPI

Review

A Review of Abnormal Crowd Behavior Recognition Technology Based on Computer Vision

Rongyong Zhao [1], Feng Hua [1,*], Bingyu Wei [1], Cuiling Li [1], Yulong Ma [1], Eric S. W. Wong [2] and Fengnian Liu [3]

[1] School of Electronic and Information Engineering, Tongji University, Shanghai 201804, China; zhaorongyong@tongji.edu.cn (R.Z.); 2331988@tongji.edu.cn (B.W.); licuiling@tongji.edu.cn (C.L.); evanma@tongji.edu.cn (Y.M.)
[2] Hongkong Institute of Water and Sainitation Safety, Hong Kong 999077, China; sw.wong@connect.polyu.hk
[3] Shanghai Karon Eco-Valve Manufacturing Co., Ltd., Shanghai 201804, China; fengnian.liu@karon-valve.com
* Correspondence: 2341070@tongji.edu.cn

Abstract: Abnormal crowd behavior recognition is one of the research hotspots in computer vision. Its goal is to use computer vision technology and abnormal behavior detection models to accurately perceive, predict, and intervene in potential abnormal behaviors of the crowd and monitor the status of the crowd system in public places in real time, to effectively prevent and deal with public security risks and ensure public life safety and social order. To this end, focusing on the abnormal crowd behavior recognition technology in the computer vision system, a systematic review study of its theory and cutting-edge technology is conducted. First, the crowd level and abnormal behaviors in public places are defined, and the challenges faced by abnormal crowd behavior recognition are expounded. Then, from the dimensions based on traditional methods and based on deep learning, the mainstream technologies of abnormal behavior recognition are discussed, and the design ideas, advantages, and limitations of various methods are analyzed. Next, the mainstream software tools are introduced to provide a comprehensive reference for the technical framework. Secondly, typical abnormal behavior datasets at home and abroad are sorted out, and the characteristics of these datasets are compared in detail from multiple perspectives such as scale, characteristics, and uses, and the performance indicators of different algorithms on the datasets are compared and analyzed. Finally, the full text is summarized and the future development direction of abnormal crowd behavior recognition technology is prospected.

Keywords: computer vision; abnormal behavior recognition; crowd abnormal behavior; deep learning; neural network; self-attention mechanism

check for updates

Citation: Zhao, R.; Hua, F.; Wei, B.; Li, C.; Ma, Y.; Wong, E.S.W.; Liu, F. A Review of Abnormal Crowd Behavior Recognition Technology Based on Computer Vision. *Appl. Sci.* **2024**, *14*, 9758. https://doi.org/10.3390/app14219758

Academic Editor: Antonio Pagliaro

Received: 22 September 2024
Revised: 11 October 2024
Accepted: 14 October 2024
Published: 25 October 2024

Copyright: © 2024 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 the acceleration of urbanization in China and the improvement in safety requirements in public places, the importance of abnormal behavior recognition in dense crowds has become increasingly prominent. At the same time, various large-scale public activities and large-scale crowd gatherings show a continuous growth trend, including various exhibitions, entertainment performances, and sports events. In these scenarios, the flow of people surges, the personnel is highly dense, and the composition is diverse and complex. Once an abnormal event or emergency occurs, it is very easy to cause panic within the crowd, which in turn leads to situations such as rapid movement of the crowd, crowded collisions, and even chaotic pushing and shoving. In this case, the crowd will fall into an unstable state and even serious stampede accidents may occur, causing a tragedy of a large number of casualties [1].

In the early computer vision tasks, artificial feature descriptors played an important role, especially in the recognition of abnormal behaviors in crowds [2]. Through artificial feature descriptors, key attributes with discrimination and invariance are extracted from image data to represent the characteristics of pedestrians and the movement state of the

crowd. However, when dealing with images with large type differences or processing complex visual scenes (such as complex backgrounds, target occlusion, and drastic changes in illumination), traditional manual feature methods can usually only capture low-level local texture and shape information in the image and cannot maintain stability and effectiveness, resulting in the performance of most abnormal detection methods being restricted by the above factors. Compared with traditional recognition methods, advanced computer vision technologies such as deep learning can automatically extract image features and better model time series data, capture the temporal and spatial characteristics of the dynamic changes in the crowd, and automatically distinguish abnormal and normal behaviors of the crowd in video frames.

In recent years, automatic recognition of abnormal behaviors in crowds based on computer vision has become a research hotspot. The literature [3] uses the FCM clustering algorithm to group the key points of the trajectory and construct a feature histogram of the cluster group motion pattern, visualizing the motion pattern features in the form of high-dimensional coordinates. The literature [4] proposes an abnormal crowd behavior detection algorithm, SFCNN-ABD, based on the streamlined flow Convolutional Neural Network. The literature [5] detects possible group abnormal events, such as panic, pushing, gathering, and other dangerous behaviors by real-time analysis of the characteristics of pedestrian movement, density changes, and interaction behaviors in the surveillance video. The literature [6] details the video-based human abnormal behavior recognition and detection technology, covering many aspects such as the classification of abnormal behaviors, feature extraction methods, and discrimination methods. The literature [7] reviews the non-invasive human fall detection system based on deep learning and elaborates on the performance indicators of models such as CNN, Auto-Encoder, LSTM, and MLP in detail. The literature [8] combs the research progress in the field of abnormal detection of surveillance videos, including many aspects such as abnormal classification, detection methods, feature representation, model construction, and evaluation criteria. The literature [3–8] represents the progress of different stages in the field of abnormal behavior recognition of crowds.

It is worth noting that there are three deficiencies in the existing literature research. (1) The analysis of abnormal behaviors in the literature is not comprehensive, and piece of the each literature only focuses on the research results of a specific development stage; (2) The pedestrian abnormal behavior detection technology has not been systematically included, and the key factors such as the characteristics and limitations of various methods have not been compared in detail; (3) For the visual problems that are common in complex scenes, such as dense crowds and severe occlusion between individuals, little of the existing literature has proposed targeted and efficient visual occlusion resolution strategies.

Therefore, given the above deficiencies, this paper conducts a comprehensive and systematic analysis of the automatic recognition technology of abnormal behaviors of crowds in the field of computer vision, so that researchers in this field can better grasp the current situation and development direction of abnormal behavior detection of crowds. The main contributions of this paper are summarized as follows:

(1) Systematic summary. This study summarizes the traditional methods of abnormal behavior detection systematically. Then we deeply explore how to use mathematical models to capture and analyze key characteristics such as the speed, direction, abnormal movement of individuals and the spatial layout, flow of people, flow speed, and direction of the group;

(2) Comprehensive method categorization. Starting from the design concept and focus of the algorithm core, the abnormal behavior detection methods based on deep learning are divided into five categories. In addition, for the scale change and occlusion problems in the crowd, this paper provides a novel dynamic pedestrian centroid model and a visual occlusion resolution strategy;

(3) Technical tools recommendation. We recommend four types of representative software tools and provide an experimental platform and practical tool reference for researchers in related fields;

(4) Character analysis of bench-mark datasets. We analyze four international bench-mark datasets of typical abnormal behaviors of crowds. Then we compare these datasets in detail from multiple perspectives such as scale, annotation, and main uses; and prospect the future development direction of abnormal behavior recognition from five aspects such as multi-modal fusion, multi-source multi-dimensional data fusion, and metaverse evolution model.

The rest of this paper is structured as follows: Section 2 gives the definitions of crowd levels and abnormal behaviors; Section 3 introduces the main challenges faced by the abnormal behavior recognition task in the dense crowd scene; Section 4 comprehensively summarizes the methods of abnormal behavior recognition from the two dimensions of traditional methods and deep learning and further introduces the current mainstream software tools; and Section 5 introduces the datasets widely used in the field of abnormal behavior detection at home and abroad and the performance indicators of each algorithm on these datasets. Section 6 summarizes this paper and presents the future development trend of this research field. The article framework is shown in Figure 1.

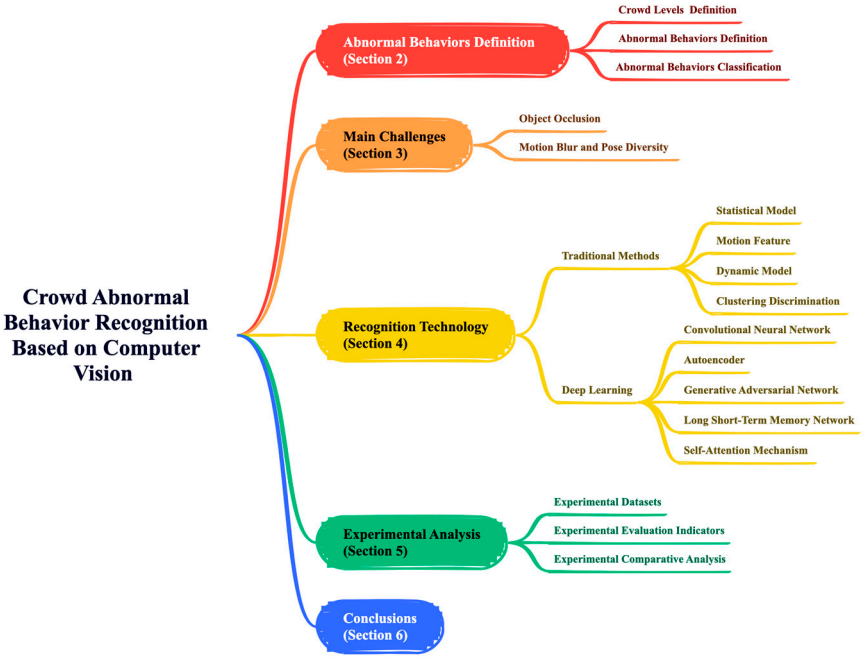


Figure 1. Paper structure diagram.

2. Abnormal Behaviors of Pedestrians in Public Places

2.1. Classification of Crowd Levels

High-density crowds usually refer to the dense concentration of pedestrians in a specific space or area, such that the distance between individuals is small, which may lead to restricted movement, blocked sight, or poor air circulation. In this case, the density of the crowd has reached a level that may affect public safety, increasing the risks of stampede, congestion, and difficult evacuation. Specifically, the quantification of crowding can be defined by "crowd density" [9], that is, the number of pedestrians in a unit area, often expressed as (p/m^2).

According to the crowd density value ranges, the stability state of the crowd can be divided into five different levels [10], as shown in Table 1: Very Low (VL), Low (L), Medium (M), High (H), and Very High (VH). Among them, VL and L are stable states, M is a critical stable state, and H and VH are unstable states. When the crowd is in a critical stable state, the safety management department should pay attention to the movement of the crowd. Once the crowd reaches an unstable state, all departments should take emergency safety management measures, such as restricting the flow of people, increasing protective fences, and dispatching more on-site safety management personnel.

Table 1. Crowd states.

Crowd State	Density Range (p/m^2)	Crowd Level
Free Flow	<0.5	Very Low (VL)
Restricted Flow	0.50–0.80	Low (L)
Dense Flow	0.81–1.26	Medium (M)
Very Dense Flow	1.27–2.00	High (H)
Congestion	>2.00	Very High (VH)

2.2. Definition and Classification of Abnormal Behaviors

2.2.1. Definition of Abnormal Behaviors

Abnormal behaviors refer to the behaviors that individuals or groups exhibit that are significantly different from the normal social behavior patterns in crowded scenes. These behaviors not only include direct physical conflicts and dangerous actions, but also involve any atypical behaviors that may lead to public disorder, the spread of panic emotions, and damage to the safety of individuals or collectives. In a high-density environment, the combined effects of the density, mobility, emotional state of the crowd, and environmental factors make the identification and management of abnormal behaviors extremely complex and urgent. The ubiquitous characteristics of abnormal behaviors [11] include ① suddenness, abnormal behaviors of pedestrians often occur suddenly without warning; ② harmfulness, causing harm to the pedestrians themselves and triggering abnormal reactions of the surrounding crowd; ③ abnormal emotions [12], when the crowd faces a certain emergency, panic emotions arise; and ④ spatial abnormality: irregular flow and aggregation, which may cause safety accidents such as stampede.

2.2.2. Classification of Abnormal Behaviors

The abnormal behaviors of the crowd can be divided into two major categories: violent and non-violent [13]. Violent abnormal behaviors usually lead to more serious casualties or property losses. Among them, common violent abnormal behaviors include group fighting, stampede, terrorist attacks, riots, panic, escape, etc. While non-violent abnormal events do not directly cause violent injuries, they may still bring a series of impacts, such as parades and onlookers.

3. Challenges Brought by Crowd Density

In high-density crowd scenes, crowd behavior recognition faces a series of challenges, which stem from the high complexity brought by the dense distribution of the crowd. Specifically, the frequent occlusion between individuals and the motion blur induced by rapid movement together constitute various obstacles in the recognition and analysis process. These factors are intertwined, greatly increasing the difficulty of accurately detecting, continuously tracking and deeply understanding crowd behaviors, and posing a severe test for realizing crowd density distribution, pedestrian trajectories, and abnormal behavior recognition [14].

3.1. Target Occlusion Problem

In a crowded crowd, the occlusion situation between individuals is complex and changeable, including not only partial occlusion, but also complete occlusion, and even the alternating occurrence of continuous occlusion and instantaneous occlusion [15], especially in public places such as railway stations, subways, and popular squares. Occlusion causes the key features of the target to be hidden or changed, making it difficult for feature-based detection and recognition algorithms to extract complete and reliable visual information. For behavior recognition models, occlusion may cause the established target feature model to fail, and it is difficult to correctly associate the targets before and after occlusion, thereby causing target loss or incorrect association. In crowded scenes, the mutual occlusion between targets is often accompanied by dense interactions between targets, such as collisions and passing by, and these interactions further increase the complexity of the

occlusion problem. The algorithm needs to consider the interaction between targets and potential group behavior patterns [16] while dealing with occlusion.

3.2. Motion Blur and Postural Diversity

The rapid change in individual movement in the crowd not only causes the target edge captured in the image to be blurred, affecting the accuracy of feature extraction based on clear contours and details; but, it may also cause misrecognition. The blur effect can confuse the understanding of target types and behaviors in visual algorithms [17]. Especially for those individuals who accelerate instantaneously or suddenly turn; their performance in the image is often trailing and residual images, which further reduces the stability and accuracy of the behavior recognition model. In addition, the joints of the human body are highly flexible, and the combination of different bending degrees and directions of each joint will show completely different postures. In the context of movement, this flexibility makes the human posture extremely susceptible to changes due to external forces and other external conditions [18]. A series of actions, whether it is head tilting, arm waving, switching between standing and walking, or adjusting the walking speed, will affect the judgment of pedestrian abnormal behaviors by the behavior recognition model.

3.3. Recommended Parameters for Hardware Resources

In high-density crowd scenarios, achieving efficient crowd behavior recognition not only faces technical challenges but also comes with the performance requirements of hardware resources. From a hardware perspective, higher resolution can provide more accurate classification and can capture larger areas and more objects, thereby improving recognition accuracy. However, high-resolution images increase the computer resources required for real-time calculation algorithms and place higher demands on processing capabilities. Therefore, combined with the current industry trends and actual needs, the best practices for abnormal behavior recognition systems are provided. Table 2 summarizes the recommended parameter configurations of cameras in specific areas such as train station waiting rooms, subway platforms, popular shopping plazas, and sports venue entrances, aiming to ensure that overall costs are controlled to meet the performance requirements.

Table 2. The recommended parameters for hardware resources.

Region Type	Crowd Sample	Sensitivity Level	Recommended IPC Resolution	Lens Requirement	Frame Rate (fps)	Recommended GPU	Remark
Train station waiting room	500–1000	High	3840 × 2160	Wide-angle to standard focal length	25–30	Gigabyte GeForce RTX3060(From China, Be purchased from JD.com)	High human flow density, requiring a high level of sensitivity to ensure safety
Subway platform	100–200	Middle	2560 × 1440	Standard focal length	25–30	Gigabyte GeForce GTX1080(From China, Be purchased from JD.com)	Medium to high human flow density, needing to balance false positives and false negatives
Popular shopping plaza	800–1200	High	3840 × 2160	Wide-angle focal length	20–25	Gigabyte GeForce RTX3060(From China, Be purchased from JD.com)	Large open space, allowing a certain range of errors
Entrance of sports venue	200–500	Middle	2560 × 1440 1920 × 1080	Standard to wide-angle focal length Infrared night vision	25–30	Gigabyte GeForce RTX2070(From China, Be purchased from JD.com)	During special events, there may be an extremely high density of people gathering

4. Research Status of Abnormal Crowd Behavior Recognition

In recent years, traditional learning methods have accumulated rich research results in the field of abnormal crowd behavior recognition. However, with the increasing complexity of crowd scenes, the recognition performance of traditional methods has certain limitations and it is often difficult to capture the subtle differences and dynamic changes in abnormal

behaviors. Against this background, the development of deep learning technology and the method of integrating neural networks for efficient pedestrian abnormal behavior recognition have gradually become research hotspots. Depending on whether neural network elements are included in the model construction, the existing abnormal behavior recognition technologies can be divided into two main classifications [19]: traditional methods-based and deep learning-based methods. The recognition technologies based on traditional methods can be classified into 4 categories: methods based on statistical models [20], based on motion features [21], based on dynamic models [22], and based on clustering discrimination [23]. And the deep learning-based methods roughly include 5 categories: methods based on Convolutional Neural Networks (CNNs) [24], based on autoencoders (AE) [25], based on generative adversarial networks (GAN) [26], based on Long Short-Term Memory Networks (LSTM) [27] and based on the self-attention mechanism (Self-Attention) [28].

4.1. Traditional Methods

This subsection classifies traditional abnormal behavior recognition methods into the following categories according to the detection principle: abnormal behavior recognition methods based on statistical models, based on motion features, based on dynamic models, and based on clustering discrimination. Most traditional abnormal behavior recognition methods first construct a normal behavior model through visual feature extraction and motion pattern analysis and then perform abnormal behavior recognition by calculating the statistics of features and setting thresholds.

4.1.1. Methods Based on Statistical Models

The core of the method based on statistical models lies in the in-depth analysis of the intrinsic distribution law of data by applying statistical theories and methods to accurately identify behaviors that deviate significantly from the normal behavior pattern. It builds statistical probability models to describe the statistical characteristics of normal behavior features and then uses these models to detect anomalies in the data.

The Gaussian Mixture Model (GMM) [29] is based on the Gaussian probability density function. By calculating parameters such as the mean and variance of pixel intensity, a model can be established, which can effectively distinguish the foreground and background and then identify the crowd behavior pattern. In high-density crowd scenes, GMM can be used to capture and describe different dynamic characteristics of the crowd, such as density, movement speed, trajectory, etc., thereby identifying abnormal behaviors that do not conform to the normal behavior pattern. The parameter estimation of GMM usually adopts the Expectation Maximization (EM) algorithm, which is an iterative algorithm including the E step (calculating the expected value) and the M step (maximizing the likelihood function) and continuously optimizes to approximate the true distribution of the data. Afig et al. [30] discussed and summarized four enhanced methods based on GMM, including the basic GMM, the combination of GMM and Markov Random Field (MRF) (GMM-MRF) [31], the Gaussian–Poisson Mixture Model (GPMM) [32], and the combination of GMM and a Support Vector Machine (SVM) (GMM-SVM) [33], pointing out that when dealing with complex crowd scenes, a combination of multiple methods can be used to improve the ability to recognize abnormal behaviors.

4.1.2. Methods Based on Motion Features

The methods based on motion features focus on analyzing the motion features of individuals or groups in the video, such as speed, direction, acceleration, and trajectory, to identify abnormal behaviors.

Among them, the motion behavior of the crowd can be effectively captured by the optical flow that changes continuously over time. Optical flow [34] is a vector field that describes the motion information of objects in an image sequence, and it reflects the position change in each pixel point in the scene between adjacent frames. In high-density crowd

scenes, the optical flow field can reveal the movement direction and speed distribution of individuals or groups, thereby identifying abnormal movements contrary to the normal behavior pattern, such as reverse movement and fast running.

Traditional motion description techniques are mostly based on the velocity information of optical flow, but the acceleration information often contains more abundant motion details. Especially when describing complex motion patterns, it can provide the information missed by the velocity descriptor and help to better understand the motion pattern. Wang et al. [35] studied an acceleration feature descriptor to improve the accuracy of abnormal behavior detection in videos. The process is shown in Figure 2. First, the optical flow field information of each frame is extracted from the video. The inter-frame acceleration is calculated through the optical flow information of two consecutive frames. The acceleration histogram feature is constructed to form the acceleration descriptor (HAVA). Using the video training set containing only normal behaviors, the normal motion pattern is learned through the energy-optimized Restricted Boltzmann Machine (RBM) model [36]. In addition, Jiang et al. [37] proposed a motion descriptor of stripe flow acceleration (SFA), and the process is shown in Figure 3. Stripe flow is a motion representation method that can effectively capture long-term spatiotemporal changes in crowded scenes. Compared with traditional optical flow, it performs better in terms of accuracy and robustness. By introducing stripe flow acceleration to explicitly model motion information, the spatiotemporal changes in crowded scenes can be accurately represented.

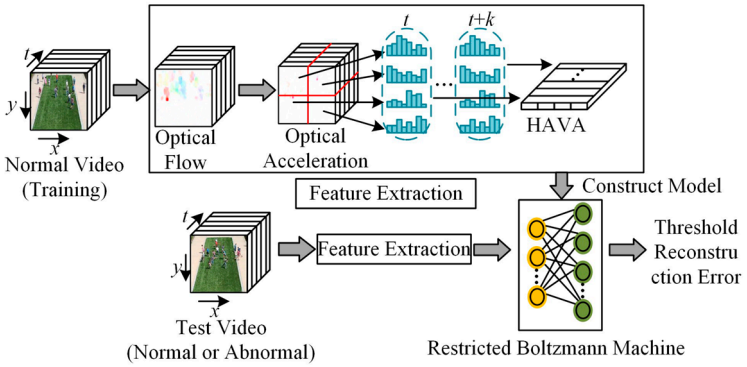


Figure 2. The framework of the acceleration optical flow method. Note, blue color indicates the period from t to $t + k$ and is used to extract the acceleration descriptors of the crowd. Yellow represents the visible layer of the restricted Boltzmann machine, and green represents the hidden layer of the restricted Boltzmann machine, which is used for behavior feature learning.

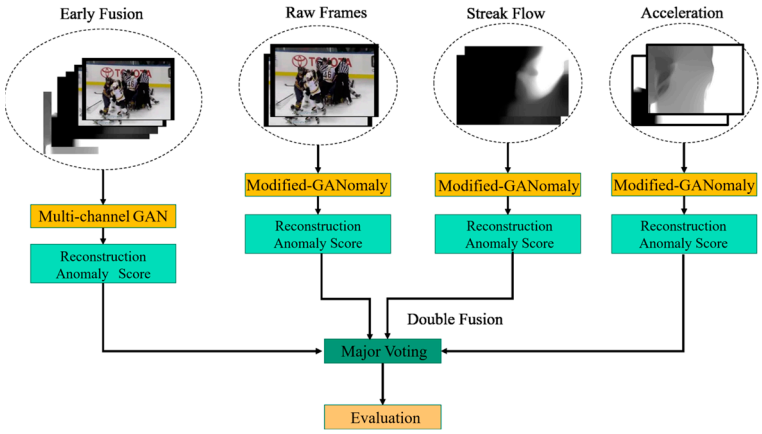


Figure 3. The framework of the SFA method.

4.1.3. Methods Based on Dynamic Models

Methods based on dynamic models, such as cellular automata models [38], particle system models [39], etc., predict the overall movement trend of the crowd by simulating the interaction between individuals and environmental constraints. These models can be used to establish a baseline behavior model for the movement of normal crowds. When the observed behavior significantly deviates from the model prediction, it can be regarded as abnormal. By adjusting model parameters, such as attractive force and repulsive force, it is possible to adapt to the dynamic characteristics of the crowd in different scenarios and improve the accuracy of abnormal behavior recognition.

Chang et al. [40] proposed a hybrid model of cellular automata and agents, dividing the space into multiple discrete units (cells), with each cell representing an individual or a small part of the crowd in the crowd, and dynamically updating its state according to preset rules to simulate the dynamic behavior of the crowd in normal or emergencies, such as movement, aggregation, evacuation, etc. They also combined it with the Agent model to further refine individual behaviors, considering individual differences such as gender, age, physical conditions, and psychological factors such as panic and conformity on evacuation behaviors. This model can identify which individual behaviors deviate from the norm, that is, abnormal behaviors.

4.1.4. Methods Based on Clustering Discrimination

Abnormal behavior recognition technology based on clustering is an unsupervised learning method. The core idea is to group behavior data into several clusters, most of which represent normal behavior patterns, and behaviors deviating from these clusters are regarded as abnormal.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that can identify clusters of any shape and can effectively mark noise points. For high-density crowd scenes, appropriate ε (neighborhood radius) and MinPts (minimum number of neighborhood points) parameters are selected, where ε determines the definition of density and MinPts determines the minimum number of points required to form a dense area. Then, DBSCAN clustering analysis is used to identify dense crowd areas under normal behavior patterns and detect abnormal points that cannot be assigned to any cluster. Chebi et al. [41] proposed a method combining DBSCAN and neural networks for the dynamic detection of abnormal crowd behaviors in video analysis. This method uses the density-based feature of the DBSCAN algorithm to identify the dense areas of the crowd in video surveillance and further analyzes the behavior patterns of these areas through neural networks to identify abnormal behaviors that are different from normal behavior patterns.

In summary, Table 3 summarizes the design idea, advantages, and disadvantages of the above various abnormal behavior recognition technologies based on traditional methods. Although these traditional methods can effectively identify abnormal behaviors to a certain extent, a common limitation is that they still rely on manually designed features and artificially set thresholds, with limited adaptability to complex and changeable scenarios, and can be vulnerable to interference from environmental factors. In addition, traditional methods cannot usually automatically learn deep-seated behavior patterns, resulting in limited generalization ability prone to false positives and false negatives. Therefore, there is an urgent need to seek new methods, which should have a strong automatic feature learning ability, adapt well to complex patterns, and capture subtle differences in crowd behaviors more accurately.

Table 3. The classification and characteristics of traditional behavior anomaly recognition methods.

Method Category	Design Idea	Advantage	Limitation	Reference
Statistical Model	Build the background model through statistical methods	Strong pattern recognition ability and capable of handling the complex dynamics of time series data	Complex parameter estimation and sensitivity to initial values	[29–33]
Motion Feature	Identify abnormalities by analyzing the motion characteristics (such as speed, acceleration, etc.) of individuals or groups	Has high sensitivity and specificity for the recognition of motion patterns	Vulnerable to environmental factors interference (such as illumination changes, occlusion)	[34–37]
Dynamic Model	Build a microscopic model of group behavior by simulating the interaction between individuals and environmental constraints	Suitable for simulation and prediction, and model parameters can be adjusted to adapt to different scenarios	High model complexity and unable to fully simulate	[38–40]
Clustering Discrimination	Group the behavior data into different clusters and identify abnormalities based on the differences between clusters	Unsupervised learning is applicable in the case of lacking labeled data	Poor processing ability for time series data	[41]

4.2. Deep Learning-Based Methods

In recent years, deep learning has become the core driving force in the field of computer vision. Compared with traditional methods, deep learning mostly uses the powerful representational learning ability of deep neural networks to automatically extract high-level abstract features from raw data, which can effectively distinguish normal behaviors from abnormal behaviors. According to the network structure and flow processing method, deep learning abnormal behavior recognition methods are classified into the following categories: abnormal behavior recognition methods based on Convolutional Neural Networks, based on autoencoders, based on generative adversarial networks, based on Long Short-Term Memory Networks, and based on self-attention mechanisms.

4.2.1. Methods Based on the Convolutional Neural Network

The Convolutional Neural Network (CNN), due to its strong ability to process pixel data and efficient learning feature representation, has become an indispensable part of the field of computer vision, such as image classification, semantic segmentation, and target feature analysis. As a deep feed-forward neural network system, the operation process of CNN can be divided into two major steps: feature extraction and classification decision-making, and the structure is shown in Figure 4. In the field of abnormal crowd behavior recognition, the current CNN-based methods mainly focus on two core directions: spatial feature and spatiotemporal feature learning.

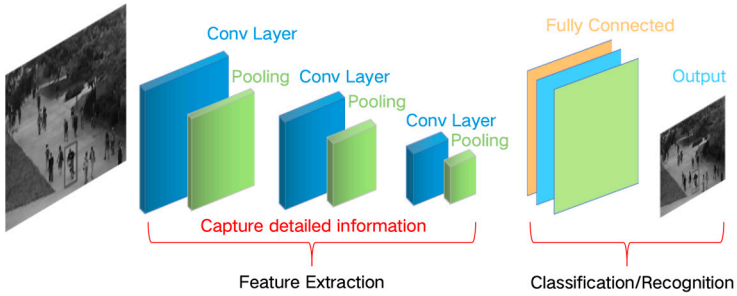


Figure 4. The structure of CNN.

(1) Spatial feature. It focuses on extracting the appearance information of pedestrians from video frames to identify atypical behavior patterns in appearance, that is, appearance anomalies. Based on different video forms and annotations, abnormal crowd behavior recognition methods can be divided into three learning methods: supervised, semi-supervised, and unsupervised.

Supervised abnormal behavior recognition is a strategy based on annotated datasets. It extracts pedestrian features through CNN and builds a classification model to identify abnormal patterns in the original data based on normal and abnormal behavior labels [42]. Since explicit category labels are provided during training, CNN can learn precise feature representations for specific tasks, thereby achieving a high accuracy rate for labeled abnormal behavior recognition. However, this method highly depends on the annotated datasets of abnormal behaviors, and the cumbersome manual annotation limits the development of supervised algorithms.

Unsupervised/weakly supervised learning methods can use CNN to extract and analyze pedestrian feature representations without explicit abnormal behavior category labels or with only a small number of labels. This method can detect unknown or unlabeled abnormal behavior types, adapt to a wider range of behavior patterns, and improve the generalization ability of the model in practical applications. Unsupervised/weakly supervised abnormal behavior recognition is implemented based on two stages: feature extraction and abnormal recognition. In the feature extraction stage, pedestrian appearance features in the video sequence are extracted through pre-trained CNNs, commonly including VGG [43], ResNet50 [44], AlexNet [45], GoogLeNet [46], Inception [47], etc. After extracting the appearance features of pedestrians, abnormal classification is performed through abnormal behavior recognition algorithms, commonly including a one-class classifier [48], Gaussian classifier [49], Support Vector Machine (SVM) [50], etc. Singh et al. [51] proposed a method of Aggregation of Ensembles (AOEs), which fused and fine-tuned the three networks of AlexNet, GoogLeNet, and VGG, especially adjusting the number of output nodes of the fully connected layer to match the binary classification task, and adopted a hierarchical fine-tuning strategy, only updating the learning rate of specific layers and keeping other layers unchanged, thereby avoiding training the network from scratch and improving the efficiency of abnormal behavior recognition. At the same time, each sub-ensemble is composed of a specific classifier and the combination of three CNN models, providing multiple classification outputs for each video frame. To solve the problem of low efficiency of video frame block convolution, Sabokrou et al. [52] proposed a method based on a Fully Convolutional Neural Network (FCN), extracting features of all regions from the entire video frame and performing convolution and pooling operations in parallel, which can achieve a processing speed of about 370 frames per second, meeting the requirements of real-time processing.

(2) Spatiotemporal feature learning. This method integrates dynamic information in the time series based on space. This is usually achieved by using optical flow data or upgrading to 3D convolution technology, aiming to capture the complex relationship between motion and time between video frames and identify abnormalities in the movement trajectories and speeds of pedestrians. Under this framework, according to the different time-perception feature fusion strategies, it is mainly divided into two implementation approaches: a two-stream CNN architecture and a three-dimensional Convolutional Neural Network (3D CNN).

The two-stream Convolutional Neural Network decomposes the processing of video data into a spatial stream and a temporal stream. The spatial stream focuses on the static content of the video frame, that is, the appearance feature in each frame image. The temporal stream focuses on analyzing the temporal changes and dynamic behaviors in the video sequence and usually uses optical flow images as the input. Hu et al. [53] proposed a weakly supervised learning ABDL framework. Firstly, the Faster R-CNN network is used to identify objects (such as pedestrians) in the scene; then, the large-scale optical flow histogram (HLSOF) is used to describe the behavior features of the objects and finally, the

behaviors are classified through the Multi-Instance Support Vector Machine (MISVM) to distinguish normal or abnormal. Wang et al. [54] designed a two-stream Convolutional Neural Network model (TS-CNN), as shown in Figure 5. By combining the above optical flow features and trajectory information, this network can process spatiotemporal information simultaneously and improve the accuracy of abnormal behavior detection. At the same time, the Kanade–Lucas–Tomasi (KLT) tracking algorithm is used to obtain the single-frame image of the crowd movement trajectory. This strategy enables the algorithm to better handle the occlusion problem in the crowd and further enhances the recognition of abnormal behaviors through the continuity feature of the behavior.

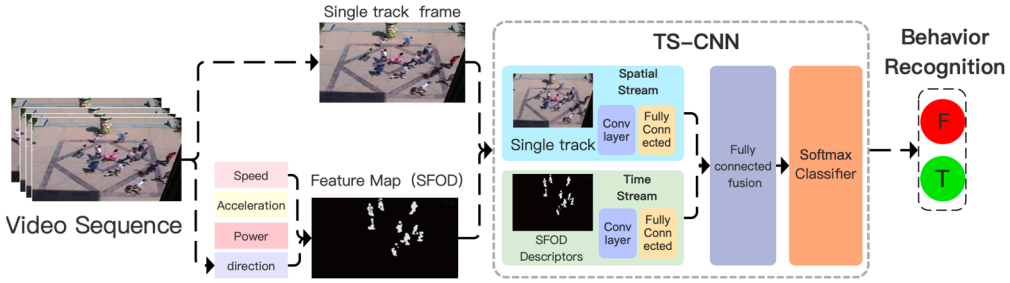


Figure 5. The structure of TS-CNN.

Three-dimensional CNN adds a time dimension based on the original 2D convolution. By performing convolution on the time axis, 3D CNN can capture the change patterns of behaviors over time, which is extremely crucial for identifying abnormal behaviors (such as falling, running, gathering, etc.) that need to consider the behavior sequence and time context. Hu et al. [55] proposed a method based on a Deep Spatiotemporal Convolutional Neural Network (DSTCNN), extending two-dimensional convolution to three-dimensional space and comprehensively considering the spatial features of static images and the temporal features between front and rear frames. Firstly, the video screen is divided into multiple sub-regions and spatiotemporal data samples are extracted from these sub-regions and input into DSTCNN for training and classification, achieving accurate detection and location of abnormal behaviors. At the same time, to solve the problems of network dispersion and insufficient recognition ability, Gong et al. [56] proposed the LDA-Net framework, as shown in Figure 6. It consists of a human body detection module and an abnormal detection module. Among them, the YOLO algorithm is introduced in the human body detection module to make the abnormal detection module more concentrated; in the abnormal detection module, 3D CNN is used to simultaneously capture the motion information of labeled normal and abnormal action sequences from the spatial and temporal dimensions.

4.2.2. Methods Based on Autoencoders

Autoencoders (AEs) [57] constitute an unsupervised learning method. The basic architecture consists of two parts, the encoder and the decoder. It reconstructs the original input by learning the effective compressed representation of the data. The structure is shown in Figure 7. In anomaly detection, AEs are trained to reconstruct normal behavior data. Thus, for those inputs significantly different from the training data (i.e., abnormal behaviors), their reconstruction errors will increase. During this process, the Mean Square Error (MSE) [58] is commonly used as the loss function. MSE is a statistical indicator that measures the difference between the predicted value and the true value, and the expression is shown in Equation (1), as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (1)$$

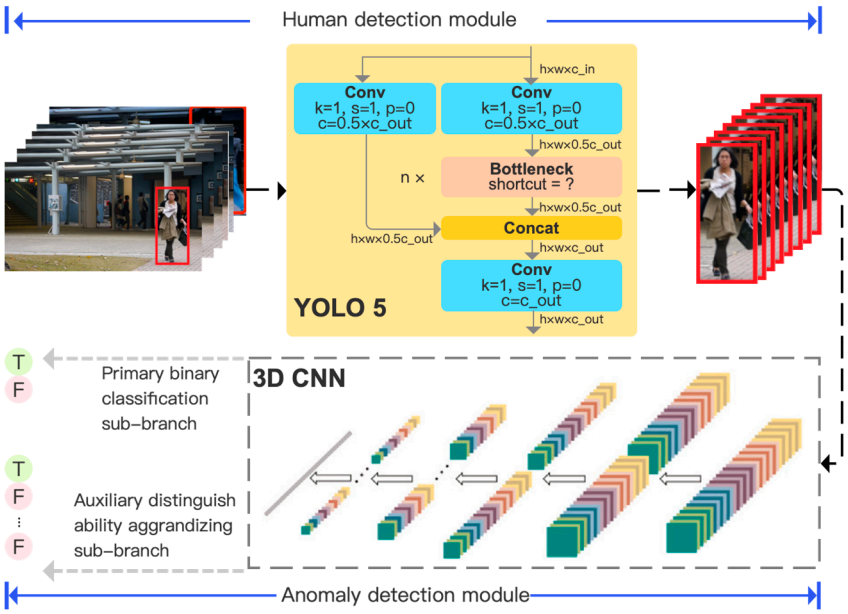


Figure 6. The structure of LDA-Net. Note, the red rectangles are the detected boundary of the moving objects. The yellow rectangle is the YOLO5 model, in which the blue ones are convolutional layers and are used for extracting features, the pink ones are bottleneck structure modules and are used for reducing network parameters, and the orange ones are Concat modules and are used for connecting feature maps of different layers.

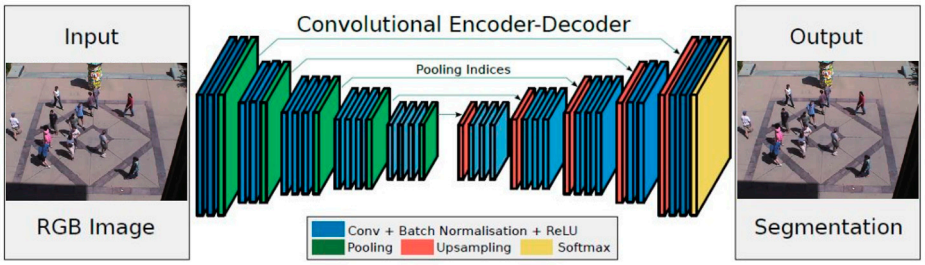


Figure 7. The structure of an AutoEncoder.

Among them, $\hat{y}_i$ is the predicted value of the model for the ith sample, y_i is the true value of this sample, and $\hat{y}_i - y_i$ is the residual (the difference between the predicted value and the true value). The smaller the value of MSE, the smaller the average gap between the predicted value and the true value of the model, and the higher the fitting degree of the model. In recent years, Autoencoders have been widely used for abnormal crowd behavior recognition. These methods are mainly divided into two categories [59]: methods based on similarity measurement and methods based on hidden feature representation learning.

(1) The method based on similarity measurement. This is a technique used to evaluate and quantify the degree of similarity between two data objects, samples, sets or vectors. In the field of abnormal crowd behavior recognition, abnormal behavior recognition is carried out by comparing the similarity between the original input image and the reconstructed image by the Convolutional Autoencoder (CAE), and the process is shown in Figure 8. The similarity between the original image and the reconstructed image is compared through similarity measurement techniques (such as Euclidean Distance [60], Pearson Correlation Coefficient [61], etc.). If the similarity is low, it indicates that there is abnormal behavior in the original image.

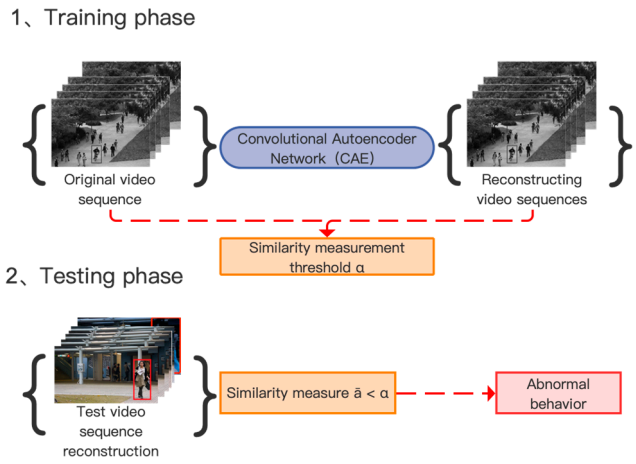


Figure 8. Anomaly behavior recognition process based on similarity measurement.

The original video sequence contains the appearance features of pedestrians, and the use of optical flow analysis can describe the dynamic trajectories of foreground targets. Information fusion can be carried out through the Convolutional Autoencoder (CAE) to extract the appearance attributes and dynamic features of pedestrians. Nguyen et al. [62] designed a CNN combining CAE and U-Net, which share the same encoder. CAE is responsible for learning the normal appearance structure, while U-Net attempts to associate these structures with the corresponding motion templates. This design aims to jointly improve the detection ability of abnormal frames through two complementary streams (one dealing with appearance and the other dealing with motion) and the model supports end-to-end training.

In addition to the fusion of optical flow analysis, some recent studies have utilized the processing ability of the Convolutional LSTM network (ConvLSTM) [63] for time series and combined it with the Convolutional Autoencoder (CAE), significantly improving the extraction ability of spatiotemporal features in videos. Xiao et al. [64] designed the Probabilistic Memory Autoencoder Network (PMAE), integrating 3D causal convolution and time dimension shared fully connected layers in the autoencoder network to extract spatiotemporal features in video frames, while ensuring the temporal sequence of information and avoiding the leakage of future information. Based on the spatiotemporal encoder, Nawaratne et al. [65] proposed the method of Incremental Spatiotemporal Learner (ISTL) for online learning of the spatiotemporal normal behavior patterns in video surveillance streams, thereby achieving abnormal detection and location. At the same time, an active learning strategy of fuzzy aggregation is introduced to dynamically adapt to unknown or emerging normal behavior patterns in surveillance videos, allowing the system to continuously refine the understanding of normal and abnormal behaviors based on environmental changes and the acquisition of new information.

(2) The method based on hidden feature representation learning. This method focuses on using an encoder to map high-dimensional input data into a low-dimensional vector space, focusing on extracting the core information of normal behavior patterns. Subsequently, the decoder attempts to reconstruct the original data. During the optimization process, it ensures that the model captures the key structure of the behavior data. The process is shown in Figure 9. Common autoencoders can be divided into Stacked Denoising Autoencoders (SDAEs) [66] and Variational Autoencoders (VAEs) [67].

SDAE extracts higher-level feature representations by combining multiple Denoising Autoencoders (DAEs) [68]. The goal of a single DAE is to reconstruct the original data on noisy data. The main principles are the following: ① the encoder maps the noisy input data to a low-dimensional hidden representation and ② the decoder reconstructs the original noise-free data based on this and minimizes the reconstruction error to learn effective information. During the stacking process, the output of the first DAE (i.e., the denoised

representation) serves as the input of the second DAE, and so on. Each layer attempts to further remove noise or extract more abstract features from the output of the previous layer. Wang et al. [69] used two SDAEs to learn the appearance features and motion features of behaviors. For the appearance feature, the input is the image block within the spatiotemporal volume around the dense trajectory; the motion feature is based on the optical flow block. Each SDAE contains multiple encoding layers, and the number of nodes is halved layer by layer until the bottleneck layer is reached. The output of this bottleneck layer is regarded as the learned deep feature. The structure is shown in Figure 10.

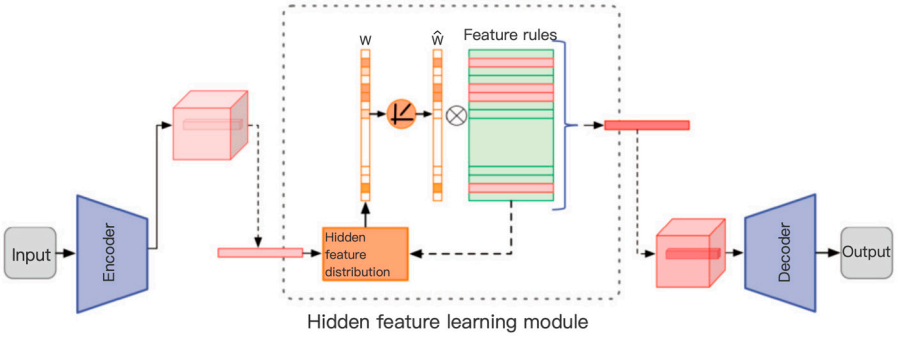


Figure 9. Abnormal behavior recognition process based on hidden feature representation learning.

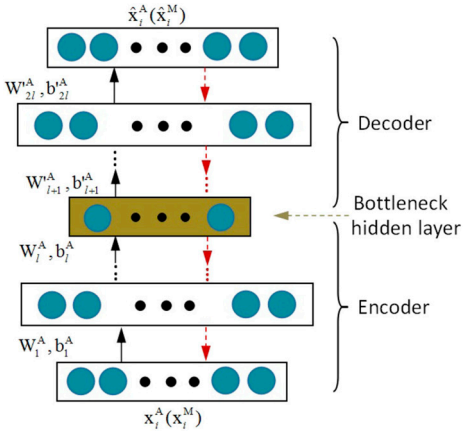


Figure 10. The structure of SDAE.

VAE is a generative model based on a probabilistic framework. Its encoder is similar to the traditional encoder, mapping the input x to a pair of vectors μ and σ, representing the mean and standard deviation of the posterior distribution $q(z \mid x)$, respectively. The decoder receives the latent variable z and attempts to reconstruct the original input x, that is, estimating $p(x \mid z)$. The output of the decoder can be regarded as the probability distribution of x given z. Wang et al. [70] designed a new method called S2-VAE. This method combines the stacked fully connected variational autoencoder (SF-VAE) and the skip convolutional variational autoencoder (SC-VAE). Among them, SF-VAE is a shallow generative network designed to simulate Gaussian mixture models to adapt to the actual data distribution. Filter obvious normal samples and improve the speed of abnormal detection. SC-VAE is a deep generative network that integrates low/middle/high-level features and reduces the loss in the information transmission process by fusing the features between the encoder and decoder layers through skip connections.

4.2.3. Methods Based on Generative Adversarial Networks

Generative adversarial networks (GAN) learn the data distribution through the adversarial process of the generator and the discriminator. In anomaly detection, a GAN can be trained to effectively generate only normal behavior data; then, the reconstruction error of

the generator or the output of the discriminator can be used to determine whether the input data are normal. Abnormal behaviors are not within the distribution range of the training data and have poor generation or reconstruction effects; thus, they are identified. The GAN model structure is shown in Figure 11. The abnormal behavior recognition strategies based on GAN can be roughly classified into two categories: the direct reconstruction or prediction error detection method and the enhanced reconstruction method combined with an autoencoder.

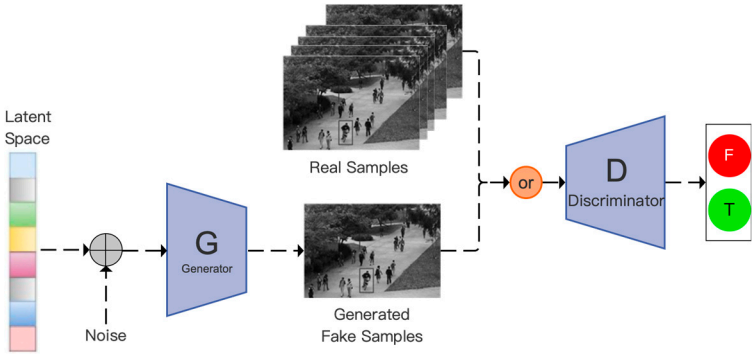


Figure 11. The structure of GAN.

The core idea of the reconstruction-based method is to train a model to learn the distribution representation of normal video data. In the testing stage, it is determined whether the test sample is abnormal based on its reconstruction error. Specifically, first, a deep learning model, that is, a neural network g, is constructed. Its goal is to learn a mapping relationship so that for any given video segment or video frame x, the output $g(x)$ processed by g is as close as possible to the original input x. Secondly, an error function f is designed, which can quantify the difference between x and $g(x)$, that is, the reconstruction error [71] $\varepsilon = f(x, g(x))$. Then, during the training process, efforts are made to find an optimal set of neural network parameters so that for all samples x in the entire dataset, the corresponding sum (or average, weighted sum, etc.) of the reconstruction errors ε reaches the minimum. Song et al. [72] proposed an Ada-Net network architecture that integrates an attention-based autoencoder with a GAN model. This architecture can adaptively learn normal behavior patterns from video data and enhance the reconstruction ability of the autoencoder through adversarial learning, making the reconstructed video frames indistinguishable from the original frames. Different from the traditional reconstruction error metric based on Euclidean distance, the researchers introduced adversarial loss for the frame discriminator to make the reconstructed frame highly similar to the original frame, thereby improving the reconstruction accuracy of the autoencoder. In addition, Chen et al. [73] proposed an anomaly detection model called NM-GAN, which consists of a reconstruction network R and a discrimination network D, forming a GAN-like architecture. NM-GAN builds an end-to-end framework, mainly consisting of three modules. (1) An image-to-image encoder–decoder reconstruction network with appropriate generalization ability. (2) A CNN-based discrimination network for identifying the spatial distribution pattern of the reconstruction error map. (3) An estimation model for quantitatively scoring anomalies. The entire model is trained in an unsupervised manner.

The prediction-based method is based on the fact that a continuous sequence of normal videos has a certain contextual dependence and regularity. Abnormal behaviors are identified by comparing the difference between the observed test frame and its predicted frame. Specifically, given the consecutive t frames of video $x_1, x_2, \ldots, x_t$, the goal of the prediction model is to generate the next frame $\hat{x}_{t+1}$ and strive to make $\hat{x}_{t+1}$ as close as possible to the actual next frame x_{t+1}. In the testing stage, it is determined as to whether the current video frame is abnormal by comparing the difference (prediction error) between the $\hat{x}_{t+1}$ predicted by the model and the actual x_{t+1}. Specifically, let h represent the prediction model, then $\hat{x}_{t+1} = h(x_1, x_2, \ldots, x_t)$. Prediction frameworks

include unidirectional prediction and bidirectional prediction: unidirectional prediction is usually based on predicting the current frame from the previous few frames. For example, Tang et al. [74] proposed the first method that combines future frame prediction and reconstruction, which is implemented through an end-to-end network. The network consists of two continuously connected U-Net modules. The first module performs future frame prediction based on the input image sequence, and the second module reconstructs the future frame based on the predicted intermediate frame, making normal and abnormal behaviors easier to distinguish in the feature space. Lee et al. [75] used the Bidirectional Multi-scale Aggregation Network (BMAN) for abnormal behavior recognition. By using bidirectional multi-scale aggregation and attention-based feature encoding, normal patterns including object scale changes and complex motions are learned. Based on the learned normal patterns, abnormal events are detected by simultaneously analyzing the appearance characteristics and motion characteristics of the scene through an appearance-motion joint detector.

The enhanced reconstruction method combining the autoencoder embeds the adversarial training logic of GAN in the training framework of AE. The autoencoder combined with GAN can not only minimize the reconstruction error but also be linked with the discriminator in GAN to output a data distribution closer to the real one. The process is shown in Figure 12. A discriminator is integrated into the architecture of the autoencoder to compare and distinguish the reconstructed image of the autoencoder network and the original input image. The quality of the reconstructed image of the autoencoder network is enhanced through adversarial training. Schlegl et al. [76] proposed a method of fast anomaly generative adversarial network (f-AnoGAN), training a Wasserstein GAN through normal samples and then training an encoder to map the image to the latent space to achieve fast inference and anomaly detection. During the anomaly detection process, the input image is reconstructed through the encoder and generator, and the combined score of the image reconstruction residual and the discriminator feature residual is used as a reliable indicator of anomaly.

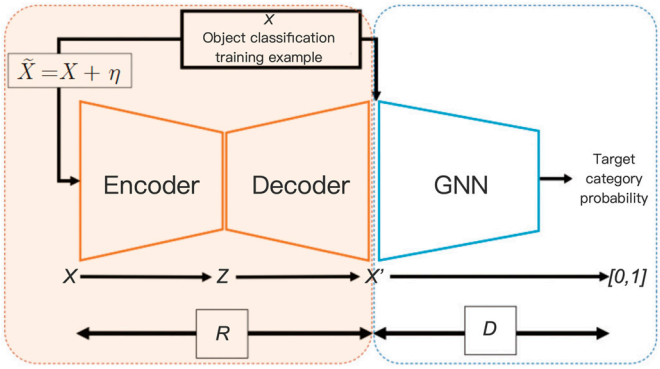


Figure 12. Enhanced reconstruction process combined with autoencoders.

4.2.4. Methods Based on Long Short-Term Memory Network

Compared with feedforward neural networks, the Recurrent Neural Network (RNN) shows significant advantages in processing sequence data and is particularly good at understanding and processing time series information. The Long Short-Term Memory (LSTM) network, as an improvement and extension of RNN, is mainly designed to solve the problems of gradient vanishing and gradient explosion encountered in the training process of long sequence data. LSTM captures and analyzes the changing patterns of behaviors over time and automatically identifies abnormal sequences that do not conform to it by learning the patterns of normal behavior sequences. LSTM introduces a gate mechanism for controlling the flow and loss of features, including three gating mechanisms: the input gate, forget gate, and output gate. The structure is shown in Figure 13.

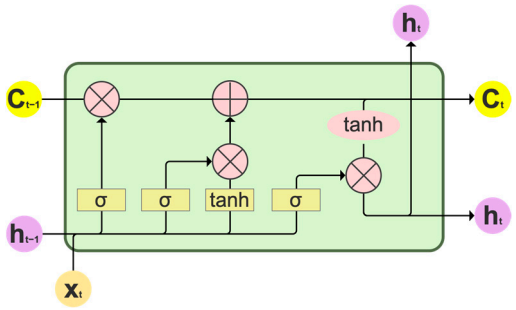


Figure 13. The structure of LSTM.

As shown in the figure above, X_t represents the input at the current moment, h_{t-1} and h_t represent the output of the previous unit and the current unit of LSTM, C_{t-1} and C_t represent the state of the previous unit and the current unit, σ represents the sigmoid activation function, and *tanh* represents the hyperbolic tangent activation function.

(1) The forget gate. This determines the information to be retained. Based on h_{t-1} and X_t, the forget gate outputs a number between 0 and 1 for the state C_{t-1}. A value of 0 means elimination and 1 means retention. The expression is shown in the Equation (2):

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{2}$$

(2) The input gate. This determines the newly incoming information. Composed of the sigmoid function and the tanh function, the result of multiplying the output values is used to update the state. The expression is shown in the Equations (3)–(5):

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\widetilde{C}_t = \tanh(W_c \bullet [h_{t-1}, x_t] + b_c)) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * C_t \tag{5}$$

(3) The output gate. This determines the output information. The output state is determined through the sigmoid layer, and the output result of the state through the tanh layer is multiplied by the output result of the sigmoid layer. The expression is shown in the Equations (6) and (7), as follows:

$$O_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{6}$$

$$H_t = O_t * \tanh(C_t) \tag{7}$$

In the field of abnormal crowd behavior recognition, researchers usually combine LSTM with other technologies to improve the recognition rate and processing speed of crowd abnormal behaviors. Meng et al. [77] integrated the focal loss function into the LSTM algorithm, optimized the model's attention to abnormal samples, and reduced the model loss. This improvement not only enhanced the sensitivity of the LSTM algorithm to abnormal behaviors but also effectively improved the detection accuracy. At the same time, the author also improved the ViBe image foreground extraction method by initializing the background model and randomly selecting the neighborhood sample set to accurately extract the foreground target, thereby simplifying the scene complexity. Wu et al. [78] combined the FCN with strong spatial feature extraction ability and the LSTM with excellent time series modeling ability. The structure is shown in Figure 14. FCN is responsible for extracting high-level semantic features from video frames, while LSTM is used to capture the changing patterns of these features over time. By simultaneously learning the spatial and temporal information in the video, it effectively distinguishes normal and abnormal behaviors in the video. Sabih et al. [79] proposed an improved end-to-end supervised learning method that combines the Long Short-Term Memory Network

(LSTM) and the Convolutional Neural Network (CNN), especially in processing optical flow features. CNN is used to extract the frame-level optical flow features calculated based on the Lucas–Kanade algorithm from the video, while the bidirectional LSTM captures the time series information of these features to jointly perform crowd abnormal detection. This method helps to understand the normal behavior patterns in the video and identify abnormal instances.

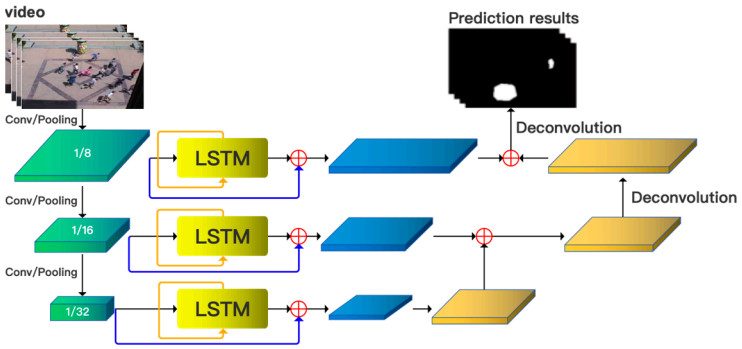


Figure 14. The structure of FCN-LSTM.

4.2.5. Methods Based on the Self-Attention Mechanism

In the field of deep learning, the attention mechanism (Self-Attention SA) [80] can be interpreted as a transformation process involving the mapping between a query vector and a series of key-value pairs, thereby generating an output vector. This output vector is obtained by performing a weighted sum of each Value vector, where the weight attached to each Value vector is determined based on the quantitative evaluation of the correlation between its corresponding Key vector and the main query vector. The calculation of the attention mechanism can be divided into three stages, as shown in Figure 15.

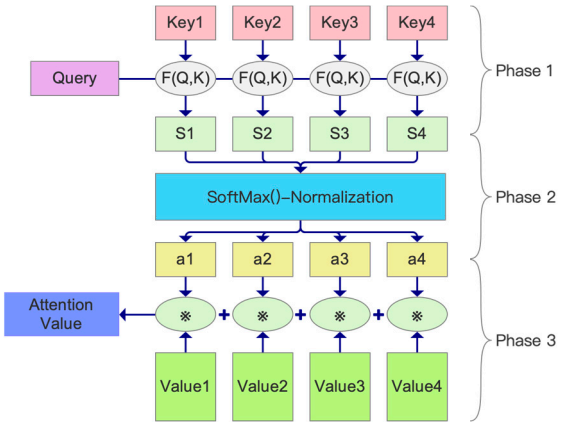


Figure 15. The process of attention mechanism computation.

Stage 1 is to calculate the similarity between Query and Key. Common methods include the vector dot product, Cosine similarity, and MLP network. The expression is shown in the Equations (8) and (9), as follows:

Vector dot product

$$Sim(Query, key_i) = Query \bullet Key_i \tag{8}$$

(1) Cosine similarity

$$Sim(Query, key_i) = \frac{Query \bullet Key_i}{\|Query\| \bullet \|Key_i\|} \tag{9}$$

(2) MLP network

$$Sim(Query, key_i) = MLP(Query \bullet Key_i) \tag{10}$$

Stage 2 introduces *SoftMax* for normalization to sort out the probability distribution of all of the elements. The expression is shown in the Equation (11), as follows:

$$a_i = SoftMax(\mathrm{Sim}_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{n} e^{Sim_i}} \tag{11}$$

Stage 3 performs weighted summation to obtain the *Attention* value. The expression is shown in Equation (12), as follows:

$$Attention(Q, K, \mathrm{V}) = SoftMax(\frac{QK^T}{\sqrt{\mathrm{d}_k}})V \tag{12}$$

Self-attention is a special form of the attention mechanism. Its uniqueness lies in that, when calculating the attention weight, the query, key, and value all come from different parts of the same input sequence. It was first widely used in the "Transformer" model, which greatly promoted the development of the field of natural language processing, and is currently also widely used in crowd sequence modeling and abnormal behavior recognition tasks. Self-attention can not only capture the long-distance dependency within the sequence but also remain efficient in parallel computing, thereby overcoming some limitations of the Recurrent Neural Network (RNN) model.

Ye et al. [81] designed a Self-Attention Feature Aggregation (SAFA) module. The structure is shown in Figure 16. This module regenerates the feature map by aggregating the embedding representations with similar information according to the attention map. At the same time, the prior bias on normal data is minimized through the self-suppression strategy, and it is used in combination with the pixel-level frame prediction error to jointly detect abnormal frames, enhancing the ability to recognize abnormalities.

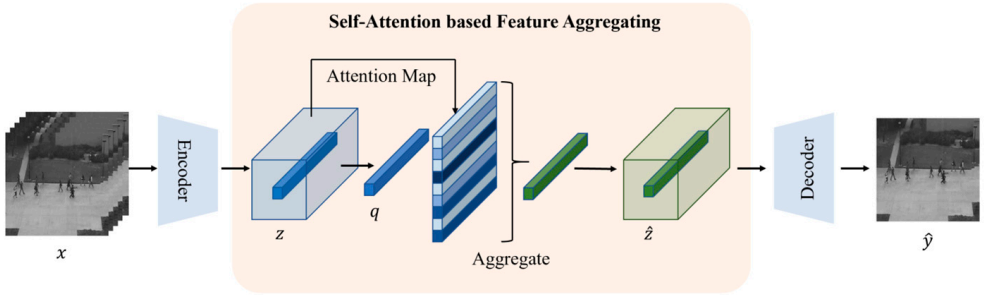


Figure 16. The structure of SAFA.

Zhang et al. [82] proposed an autoencoder model that integrates the global self-attention mechanism to capture the interaction information between the overall features of the image, facilitating the model to understand the context in which the behavior occurs and improving the accuracy of abnormal behavior determination. At the same time, a memory module is embedded in the bottleneck layer of the autoencoder to constrain the model's over-generalization of normal behaviors.

Zhang et al. [83] fused the attention mechanism and the bidirectional Long Short-Term Memory Autoencoder Network (SABiAE). The structure is shown in Figure 17. The encoder with the self-attention mechanism captures global appearance features, and the self-attention bidirectional LSTM network is used to reduce the loss of target features, thereby extracting the information between video frames globally. The model reconstructs the frame through the decoding process and determines abnormal behaviors based on the reconstruction error.

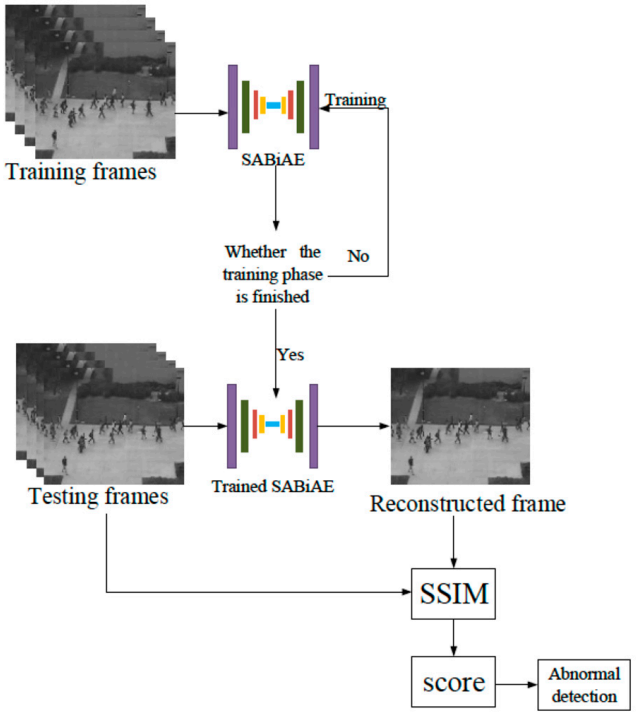


Figure 17. The structure of SABiAE. Note, SABiAE is composed of non-local encoder, non-local bidirectional long short-term memory network and decoder.

Zhang et al. [84] introduced the self-attention mechanism in the Generative Adversarial Network (GAN), using the autoencoder containing the dense residual network and the self-attention mechanism as the generator. At the same time, a new discriminator is proposed, which integrates the self-attention module based on the relative discriminator, thereby avoiding the problem of gradient vanishing during the training process and ensuring that the frames generated by the generator are closer to the real frames.

Singh et al. [85] proposed an Attention-guided Generator and Dual Discriminator Adversarial Network (A2D-GAN), the structure if which is shown in Figure 18, for real-time video anomaly detection. A2D-GAN utilizes the encoder–decoder architecture, in which the encoder adds multi-level self-attention to focus on key regions and capture context information, while the decoder uses channel attention to emphasize important features and identify abnormal patterns specific to each frame.

In addition, to solve problems such as scale changes and occlusions in the crowd, fusion models such as multi-scale feature attention fusion networks and visual occlusion resolution have emerged for abnormal crowd behavior detection. For example, Sharma et al. [86] introduced a scale-aware attention module in CNN, using the cascade of multiple self-attention branches to enhance the recognition ability for scale changes. This architecture integrates motion information, motion influence maps, and features based on the energy level distribution to achieve frame-level abnormal behavior analysis. Zhao et al. [87] proposed a Dynamic Pedestrian Centroid Model (DCM). The key points of the human skeleton are extracted from the image, the pedestrian joint sub-segments are constructed accordingly, and the dynamic characteristics of pedestrians in the video sequence are described mathematically. Experiments show that DCM can detect the U-turn behavior 277 ms earlier on average and detect the fall behavior 562 ms earlier on average to find abnormalities. At the same time, to solve the problem of partial occlusion of pedestrians, this paper also designed an anti-occlusion algorithm, as shown in the schematic in Figure 19. Firstly, cluster the key points of the pedestrian skeleton, divide the 21 key points into the head, upper body, and lower body, calculate the coordinates of the occluded part, estimate the effective position combined with the mechanical principle, and apply it in combination with DCM. This method has been verified by the fall behavior detection experiment and

the AUC on the 50 Volu. The U-turn dataset reaches 82.13%, and the AUC on the 50 Volu. The fall-down dataset reaches 86.70%.

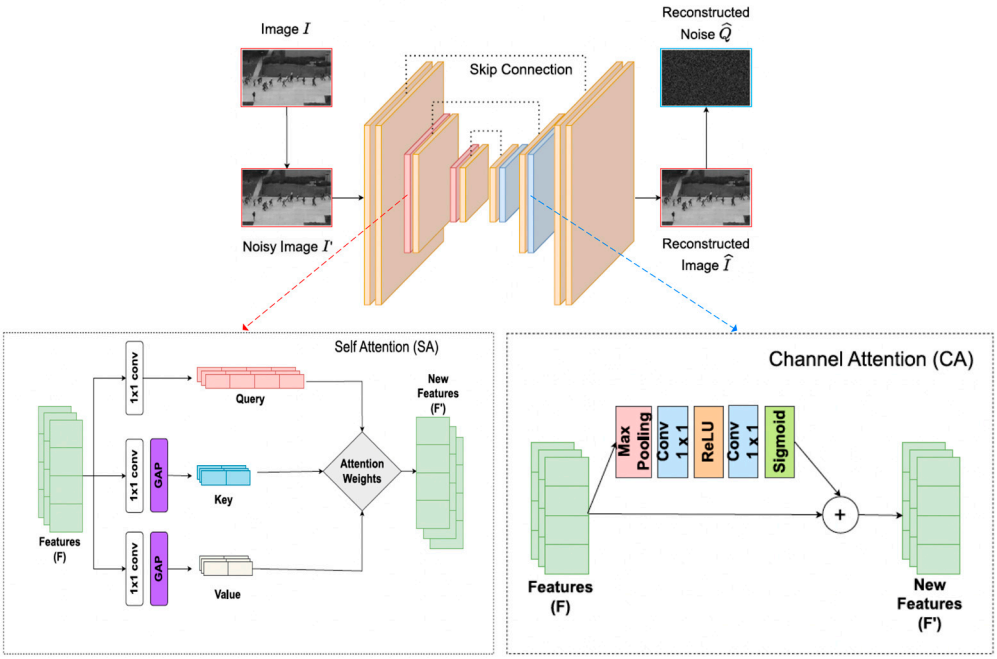


Figure 18. The structure of A2D-GAN.

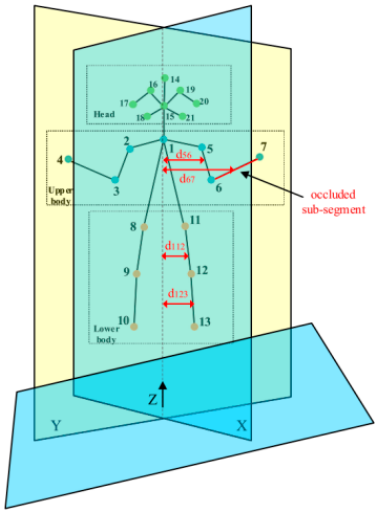


Figure 19. Diagram of the algorithm for occlusion removal based on skeleton points.

In summary, Table 4 summarizes the design ideas, advantages, and disadvantages of the above various deep learning-based abnormal behavior recognition technologies. Overall, these deep learning-based methods benefit from the powerful learning ability of neural networks and can be applied to the recognition of abnormal behaviors of pedestrians in various crowd scenarios. At the same time, a strategy of using multiple methods in combination can be adopted to further improve the generalization ability and recognition accuracy of the model.

Table 4. The classification and characteristics of deep learning behavior anomaly recognition methods.

Method	Design Ideas	Advantages	Limitations	References
CNN	Through local receptive fields and pooling operations, capture the spatial hierarchical features of the original data	Good at extracting local features and combining them into more complex patterns	Lack of time series processing capabilities	[42–56]
AutoEncoder	An unsupervised learning method that captures the main features of the data by encoding and decoding the original data	Reduce the data dimension and extract key features	Easily lead to overfitting of training data	[57–70]
GNN	It is composed of a generator and a discriminator, and the two play a game.	Reconstruct and generate new samples close to real data	Prone to model collapse and non-convergence of training	[71–76]
LSTM	Introduces forget gates, input gates, and output gates to solve the problems of gradient vanishing and gradient explosion	Good at processing sequence data of any length	The structure is complex, and training and reasoning take a long time	[77–79]
Self-Attention	Automatically assigns different attention weights according to different parts of the input sequence	Capture global interdependencies	Easy to overfit on small datasets	[80–87]

4.3. Mainstream Software Tools

In recent years, with the deep integration and breakthrough progress of big data and computer vision technology, the analysis of crowd abnormal behaviors has become an important research topic in the field of public safety. Algorithms and software tools in this field have also emerged continuously. Each representative software tool aims to reveal and identify changes and deviations from the normal patterns of crowd behaviors in complex scenarios from different dimensions. Many algorithm manufacturers focusing on this field have also launched a series of advanced software tools specifically for abnormal crowd behavior recognition. The following is an overview of the mainstream tools.

The PP-Human abnormal crowd behavior recognition tool (PaddleDetectionv2.7.0) is an industrial-level open-source real-time pedestrian analysis tool that integrates core capabilities such as object detection, tracking, and key point detection. It can adapt to different light conditions, complex backgrounds, and cross-lens scenarios.

The SenseTime abnormal crowd behavior recognition tool (Version 2.7.4 ©2014-2024) analyzes the overall characteristics and individual behavior characteristics of the crowd in the surveillance images, masters the activity rules of the crowd, and realizes the automation or even intelligence of video surveillance.

The Hikvision abnormal crowd behavior recognition tool (V3.12.0.7_E) provides powerful analysis data for crowd control, prevents the occurrence of crowding and stampede incidents, and provides effective guarantees for social security.

The Keda abnormal crowd behavior recognition tool (6.0.29.20230128_beta) combines the general intelligent recognition ability of the large model and the precise recognition ability of the small model to achieve group scene analysis such as holding knives, fighting, running, gathering, etc.

5. Comparative Analysis of Different Algorithm Experiments

5.1. Experimental Dataset Analysis

In the field of computer vision, due to the diversity and complexity of real abnormal behaviors in the research of abnormal crowd behavior analysis, there is a lack of high-quality datasets. To overcome this shortage and improve the performance of algorithms, researchers have constructed a series of datasets with different characteristics, focusing on parameters such as duration, size, resolution, etc., and covering various monitoring environments and scenarios, providing an important reference basis for the research of crowd anomaly recognition. Commonly used datasets mainly include UCSD, UMN, Shanghai-

Tech, CUHK-Avenue, and other related datasets. Figure 20 shows some abnormal behavior samples in these four abnormal behavior datasets.

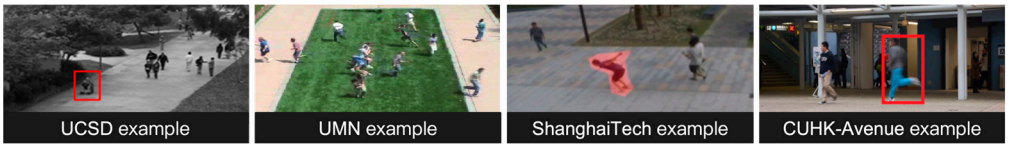


Figure 20. Examples of anomalies in the datasets. Note, the red rectangles are the detected boundary of the moving objects.

5.1.1. UCSD

The UCSD dataset [88] is a dataset for crowd behavior analysis and abnormal detection. The dataset contains two subsets, Ped1 and Ped2. The Ped1 subset includes 34 video clips, recording the pedestrian activities from the campus crosswalk scene; the Ped2 subset provides 12 video clips of the same specification, presenting similar but different pedestrian crossing area activities.

5.1.2. UMN

The UMN dataset [89] is mainly used for the evaluation and research of pedestrian detection and tracking algorithms. This dataset consists of 19 videos, including a series of pedestrian video sequences in complex backgrounds, providing various indoor and outdoor environments such as campuses, shopping malls, streets, and other pedestrian activity video clips in different backgrounds.

5.1.3. ShanghaiTech

The ShanghaiTech dataset [90] is a dataset for video abnormal detection and crowd counting. This dataset has 13 complex scenes, recording the pedestrian flow on the campus. It also contains 130 abnormal events (such as fighting, falling, running, etc.) and more than 270,000 training frames, with a duration ranging from about 30 s to 90 s.

5.1.4. CUHK-Avenue

The CUHK database [91] is a database about crowd behavior scenes. It includes crowd videos collected from many different environments with different densities and perspective ratios, such as streets, shopping centers, airports, and parks. It consists of traffic datasets and pedestrian datasets. There are a total of 474 video clips in 215 scenes in the dataset.

Table 5 summarizes and compares the above commonly used datasets of crowd abnormal behaviors from the dimensions of scene description, scale, abnormal behavior, video resolution, objects, and limitations of the dataset.

Table 5. Comparison of crowd anomaly behavior datasets.

Name	Scene Description	Scale	Resolution	Abnormal Behavior	Object	Limitation
UCSD [88]	The crowd movement on the sidewalk from the perspective of the surveillance camera	Ped1 14,000 frame 34 Training segments 36 Test segments	238 × 158	Fast moving, Reverse driving Riding a bicycle, Driving a car, Sitting in a wheelchair, skateboarding, etc.	individual	Low resolution; few types of abnormal behaviors; Relatively simple background
		Ped2 4560 frame 12 Test segments	360 × 240			

Table 5. *Cont.*

Name	Scene Description	Scale	Resolution	Abnormal Behavior	Object	Limitation
UMN [89]	Video clips of pedestrian activities in different backgrounds such as campuses, shopping malls, and streets	8010 frame 11 Video segment	320 × 240	The crowd suddenly scattered, ran, and gathered.	group	Simple background; Limited abnormal types
Shanghai-Tech [90]	13 campus area scenes with complex lighting conditions and camera angles	317,398 frame 130 Video segment	856 × 480	Crowd gathering, fighting, running, cycling.	group	Abnormal events are repetitive; The annotations have errors
CUHK-Avenue [91]	Video surveillance clips of outdoor public places	30,625 frame 16 Training segments 21 Test segments	640 × 360	Pedestrians fighting, throwing objects, running.	individual Vehicle	The shooting angle is single; The resolution is low

5.2. Experimental Evaluation Indicators

The evaluation of abnormal crowd behavior recognition integrates multiple indicators to ensure that the recognition is not only accurate but also precisely located. During the evaluation, frame-level and pixel-level standards are usually considered. Frame-level detection focuses on determining whether there is any abnormality within the video frame. Even if the specific location of the abnormality is not precise, as long as there is abnormal behavior in the frame, the entire frame is marked as abnormal; the pixel-level standard further requires the precise positioning of the spatial location of the abnormal behavior, usually requiring that the predicted abnormal pixels cover at least 40% of the real abnormal area to evaluate the accuracy of abnormal positioning.

The performance of abnormal crowd behavior recognition is usually evaluated by indicators such as the confusion matrix, ROC curve and AUC, and Equal Error Rate (EER).

By comprehensively considering the four basic indicators of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the confusion matrix is used for basic classification performance evaluation.

The ROC curve depicts the classification performance of the algorithm through the changes in True Positive Rate (TPR) and False Positive Rate (FPR). AUC, that is, the area under the ROC curve, reflects the average performance of the classifier under all thresholds. The closer the AUC value is to 1, the better the performance of the classifier.

It is defined as the error classification rate when the True Positive Rate is equal to the False Positive Rate. The smaller the EER value, the better the performance of the abnormal detection method because it balances missed alarms and false alarms and is an important indicator to measure the performance of a binary classification system.

5.3. Experimental Comparison

Table 6 summarizes the performance of traditional methods and deep learning-based abnormal behavior recognition methods on UCSD, UMN, Shanghai Tech, and Avenue datasets in recent years. This paper compares and presents evaluation indicators such as AUC and EER. These evaluation data are all cited from research papers in recent years.

Table 6. Comparison of abnormal behavior recognition experiments.

Classification	Method	Frame Level AUC/EER/%										Year
		UCSDPed1		UCSDPed2		UMN		Shanghai Tech		Avenue		
		EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	
Motion Feature	HAVA + HOF [35]	--	--	--	--	--	99.34	--	--	--	--	2020
	SFA [37]	--	--	--	96.40	--	96.55	--	--	--	--	2022
Clustering Discrimination	DBSCAN [41]	--	--	--	97.20	2.80	97.20	30.70	71.50	--	--	2020
CNN	FCN [52]	--	--	11.00	--	--	--	--	--	--	--	2019
	ABDL [53]	22.00	--	16.00	--	5.80	98.90	--	--	21.00	84.50	2020
	TS-CNN [54]	--	--	--	--	--	99.60	--	--	--	--	2020
	DSTCNN [55]	--	99.74	--	99.94	--	--	--	--	----	--	2020
	LDA-Net [56]	--	--	5.63	97.87	--	--	--	--	--	--	2020
	CAE-UNet [62]	--	--	--	96.20	--	--	--	--	--	86.90	2019
AE	PMAE [64]	--	--	--	95.90	--	--	--	72.90	--	--	2023
	ISTL [65]	29.80	75.20	8.90	91.10	--	--	--	--	29.20	76.8	2019
	S2-VAE [70]	14.30	94.25	--	--	--	99.81	--	--	--	87.6	2019
	Ada-Net [72]	11.90	90.50	11.50	90.70	--	--	--	--	17.60	89.20	2019
GAN	NM-GAN [73]	15.00	90.70	6.00	96.30	--	--	17.00	85.30	15.30	88.60	2021
	D-UNet [74]		84.70		96.30	--	--	--	73.00		85.10	2019
	BMAN [75]	--	--	--	96.60	--	99.60		76.20		90.00	2019
LSTM	FocalLoss-LSTM [77]	--	--	--	--	--	99.83	--	--	--	--	2021
	FCN-LSTM [78]	--	--	--	98.20	--	93.70	--	--	--	--	2021
	CNN-LSTM [79]	--	94.83		96.50	--	--	--	--	--	--	2022
SA	SAFA [81]	--	--	--	96.80	--	--	--	--	--	87.30	2023
	SA-AE [82]	--	--	--	95.69	--	--	--	--	--	84.10	2023
	SABiAE [83]	--	--	9.80	95.60	--	--	--	--	20.90	84.70	2022
	SA-GAN [84]	--	--	--	--	--	--	--	75.70	--	89.20	2021
	A2D-GAN [85]	9.70	94.10	5.10	97.40	--	--	25.20	74.20	9.00	91.00	2024
	SA-CNN [86]	--	--	--	--	--	99.29	--	--	--	--	2023

Note, the performance indicators of HAVA + HOF and SFA in motion features and DBSCAN in cluster discrimination in UCSD Ped2, UMN, and Shanghai Tech datasets are included to obtain a baseline in the comparison.

According to the comparison of the experimental results, the following conclusions can be drawn:

(1) The limitations of recognition technologies. Although most methods have achieved high recognition rates on some datasets, on datasets such as UCSDPed1, the EER and AUC values of many methods indicate that there are certain false alarms or missed alarms, meaning that there is still room for further optimization of abnormal behavior recognition technology, especially when dealing with specific types of data or complex scenarios;

(2) The advantages of temporal models. Two-dimensional convolutional networks that only rely on spatial features often perform worse in abnormal behavior recognition tasks than those models that integrate optical flow features, 3D convolution, and LSTM. These models incorporate temporal information and can more comprehensively capture the dynamic changes in video sequences, not only considering spatial features but also in-depth analysis of the continuity and change trends of behavior patterns in the time dimension;

(3) The advantages of Autoencoder Network. The hidden feature representation learning method based on the autoencoder network is usually superior to the similarity measurement method in terms of automatic feature learning, adaptability, and generalization ability. It is suitable for processing large-scale, high-dimensional, and complex abnormal behavior recognition tasks, especially in the unsupervised learning paradigm, and it can effectively learn the low-dimensional representation of high-dimensional data;

(4) The advantages of self-attention mechanism. The introduction of the self-attention mechanism significantly improves the accuracy and efficiency of abnormal behavior recognition. This mechanism enhances the model's ability to capture long-distance dependencies in the sequence. By allowing the model to dynamically adjust the degree of attention to different parts of the input, key features can be extracted more effectively, promoting the abnormal behavior recognition technology to a new height.

6. Conclusions

This review conducts a comprehensive and in-depth study of abnormal crowd behavior recognition technology, performing detailed analyses from four dimensions: basic definitions, traditional methods, deep learning, and performance indicators. Through a

clear classification of crowd levels, "crowd density" is quantified and divided into five levels. Next, this study analyzes recognition techniques for traditional abnormal behavior based on statistical models, motion features, dynamic models, and cluster discrimination. Although these methods are effective in specific scenarios, they still have limitations in dealing with high-dimensional, nonlinear, and complex dynamic data, such as reliance on manually designed features and limited generalization ability.

With the development of AI technologies, deep learning has become a research hotspot because of its strong automatic feature learning ability and good adaptability to complex patterns. This article discusses the methods based on the Convolutional Neural Network (CNN), Generative Adversarial Network (GAN), Long Short-Term Memory Network (LSTM), Autoencoder (AE), and Self-Attention Mechanism (SA) in detail, demonstrating their outstanding performance in abnormal behavior recognition. These methods can extract high-level features from raw data and effectively distinguish between normal and abnormal behaviors, thereby improving recognition accuracy and robustness.

In conclusion, by comparing the performance indicators of each algorithm model in Tables 4 and 6, it can be found that deep learning methods show significant advantages in dealing with complex and changeable scenarios. Although each model has its application scope and limitations, combining different methods can achieve better results. For example, algorithms such as S2-VAE, TS-CNN, and SA-CNN have achieved recognition accuracy of more than 99% in dealing with some datasets in Table 6. Therefore, in practical applications, choosing the appropriate single model or combining multiple methods based on specific requirements can effectively enhance the overall performance of the abnormal behavior recognition system and achieve more accurate and reliable detection results.

Future research can focus on improving the fusion ability of deep learning and multi-modalities, introducing context understanding and situational reasoning, and improving the robustness and adaptive learning ability of the model. It can help further broaden the recognition of multi-source and multi-dimensional data fusion, construct a self-consistent metaverse virtual–real fusion evolution model, and establish a "perception-prediction-intervention-construction" virtual–real fusion presentation mechanism. The continuous innovations in this field are expected to achieve a transformation from passive monitoring to active prevention in the future and focus on more intelligent, precise, and automated abnormal behavior recognition systems.

Author Contributions: Conceptualization, R.Z. and C.L.; methodology, R.Z. and F.H.; software, F.H.; validation, formal analysis, and investigation, R.Z., F.H. and B.W.; resources, data curation, writing—original draft preparation, writing—review and editing, and visualization, R.Z., F.H., C.L. and B.W.; supervision and data analysis support, Y.M.; current situation investigation, assessment, E.S.W.W.; review, state-of-the-art methods analysis, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 72374154).

Conflicts of Interest: Author Fengnian Liu was employed by the company Shanghai KARON ECO-VALVE Manufac-turing Co. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Haghani, M.; Lovreglio, R. Data-based tools can prevent crowd crushes. *Science* **2022**, *378*, 1060–1061. [CrossRef] [PubMed]
2. Luo, L.; Xie, S.; Yin, H.; Peng, C.; Ong, Y.S. Detecting and Quantifying Crowd-level Abnormal Behaviors in Crowd Events. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 6810–6823. [CrossRef]
3. Fadhel, M.A.; Duhaim, A.M.; Saihood, A.; Sewify, A.; Al-Hamadani, M.N.; Albahri, A.S.; Alzubaidi, L.; Gupta, A.; Mirjalili, S.; Gu, Y. Comprehensive systematic review of information fusion methods in smart cities and urban environments. *Inf. Fusion* **2024**, *107*, 102317. [CrossRef]
4. Zhang, B.; Ge, S.; Wang, Q.; Li, P. Research on behavior recognition method based on multi-order information fusion. *Acta Autom. Sin.* **2021**, *47*, 609–619. [CrossRef]

5. Jiang, J.; Zhang, Z.; Gao, M.; Xu, L.; Pan, J.; Wang, X. A crowd abnormal behavior detection algorithm based on pulse line flow convolutional neural network. *J. Eng. Sci. Technol.* **2020**, *52*, 215–222.
6. Xiao, J.; Shen, M.; Jiang, M.; Lei, J.; Bao, Z. Abnormal behavior detection in surveillance videos by fusing bag attention mechanism. *Acta Autom. Sin.* **2022**, *48*, 2951–2959. [CrossRef]
7. Alam, E.; Sufian, A.; Dutta, P.; Dutta, P.; Leo, M. Vision-based human fall detection systems using deep learning: A review. *Comput. Biol. Med.* **2022**, *146*, 105626. [CrossRef]
8. Yang, F.; Xiao, B.; Yu, Z. Anomaly Detection and Modeling of Surveillance Video. *J. Comput. Res. Dev.* **2021**, *58*, 2708–2723. [CrossRef]
9. Li, Y.C.; Jia, R.S.; Hu, Y.X.; Sun, H.M. A Weakly-Supervised Crowd Density Estimation Method Based on Two-Stage Linear Feature Calibration. *IEEE/CAA J. Autom. Sin.* **2024**, *11*, 965–981. [CrossRef]
10. Zhao, R.; Dong, D.; Wang, Y.; Li, C.; Ma, Y.; Enríquez, V.F. Image-based crowd stability analysis using improved multi-column convolutional neural network. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 5480–5489. [CrossRef]
11. Zhao, R.; Wei, B.; Zhu, W.; Zheng, C.; Li, H. Review of pedestrian abnormal behavior recognition methods in public places. *China Saf. Sci. J.* **2024**, *34*, 83–93. [CrossRef]
12. Wang, D.; Du, Y.; Dong, L.; Hu, W.; Li, B. Small sample learning algorithm based on feature transformation and metric network. *Acta Autom. Sin.* **2024**, *50*, 1305–1314.
13. Ng, S.H.; Platow, M.J. The violent turn in non-violent collective action: What happens? *Asian J. Soc. Psychol.* **2024**, *27*, 303–317. [CrossRef]
14. Chen, C.; Bai, S.; Huang, L.; Wang, X.; Liu, C. Research on passenger flow monitoring and early warning in crowded places based on video analysis. *J. Saf. Sci. Technol.* **2020**, *16*, 143–148.
15. Wang, X.; Li, Y. Edge Detection and Simulation Analysis of Multimedia Images Based on Intelligent Monitoring Robot. *Informatica* **2024**, *48*. [CrossRef]
16. Chen, X.; Yan, B.; Zhu, J.; Lu, H.; Ruan, X.; Wang, D. High-performance transformer tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 8507–8523. [CrossRef]
17. Farooq, M.U.; Saad MN, B.M.; Malik, A.S.; Salih Ali, Y.; Khan, S.D. Motion estimation of high-density crowd using fluid dynamics. *Imaging Sci. J.* **2020**, *68*, 141–155. [CrossRef]
18. Zhou, Z.; Sun, R.; Guo, C.; Zhu, Y.; Hierarchical Motion Estimation Method for Spatially Unstable Targets under Gaussian Mixture Model. Acta Optica Sinica:1-19. Available online: http://kns.cnki.net/kcms/detail/31.1252.O4.20240517.1540.029.html (accessed on 26 May 2024).
19. Zhou, Y.; Liu, C.; Ding, Y.; Yuan, D.; Yin, J.; Yang, S.H. Crowd descriptors and interpretable gathering understanding. *IEEE Trans. Multimed.* **2024**, *26*, 8651–8664. [CrossRef]
20. Alhothali, A.; Balabid, A.; Alharthi, R.; Alzahrani, B.; Alotaibi, R.; Barnawi, A. Anomalous event detection and localization in dense crowd scenes. *Multimed. Tools Appl.* **2023**, *82*, 15673–15694. [CrossRef]
21. Nayan, N.; Sahu, S.S.; Kumar, S. Detecting anomalous crowd behavior using correlation analysis of optical flow. *Signal Image Video Process.* **2019**, *13*, 1233–1241. [CrossRef]
22. Aziz, Z.; Bhatti, N.; Mahmood, H.; Zia, M. Video anomaly detection and localization based on appearance and motion models. *Multimed. Tools Appl.* **2021**, *80*, 25875–25895. [CrossRef]
23. Matkovic, F.; Ivasic-Kos, M.; Ribaric, S. A new approach to dominant motion pattern recognition at the macroscopic crowd level. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105387. [CrossRef]
24. Ganga, B.; Lata, B.T.; Venugopal, K.R. Object detection and crowd analysis using deep learning techniques: Comprehensive review and future directions. *Neurocomputing* **2024**, *597*, 127932. [CrossRef]
25. Madan, N.; Ristea, N.C.; Ionescu, R.T.; Nasrollahi, K.; Khan, F.S.; Moeslund, T.B.; Shah, M. Self-supervised masked convolutional transformer block for anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 525–542. [CrossRef]
26. Chai, L.; Liu, Y.; Liu, W.; Han, G.; He, S. CrowdGAN: Identity-free interactive crowd video generation and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2856–2871. [CrossRef]
27. Song, X.; Chen, K.; Li, X.; Sun, J.; Hou, B.; Cui, Y.; Zhang, B.; Xiong, G.; Wang, Z. Pedestrian trajectory prediction based on deep convolutional LSTM network. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 3285–3302. [CrossRef]
28. Huang, Y.; Abeliuk, A.; Morstatter, F.; Atanasov, P.; Galstyan, A. Anchor attention for hybrid crowd forecasts aggregation. *arXiv* **2020**, arXiv:2003.12447.
29. He, F.; Xiang, Y.; Zhao, X.; Wang, H. Informative scene decomposition for crowd analysis, comparison and simulation guidance. *ACM Trans. Graph. (TOG)* **2020**, *39*, 50–51. [CrossRef]
30. Afiq, A.A.; Zakariya, M.A.; Saad, M.N.; Nurfarzana, A.A.; Khir, M.H.M.; Fadzil, A.F.; Jale, A.; Gunawan, W.; Izuddin, Z.A.A.; Faizari, M. A review on classifying abnormal behavior in crowd scene. *J. Vis. Commun. Image Represent.* **2019**, *58*, 285–303. [CrossRef]
31. Goel, S.; Koundal, D.; Nijhawan, R. Learning Models in Crowd Analysis: A Review. *Arch. Comput. Methods Eng.* **2024**, 1–19. [CrossRef]
32. Rajasekaran, G.; Raja Sekar, J. Abnormal Crowd Behavior Detection Using Optimized Pyramidal Lucas-Kanade Technique. *Intell. Autom. Soft Comput.* **2023**, *35*, 2399–2412. [CrossRef]

33. Gündüz, M.Ş.; Işık, G. A new YOLO-based method for real-time crowd detection from video and performance analysis of YOLO models. *J. Real-Time Image Process.* **2023**, *20*, 5. [CrossRef] [PubMed]
34. Miao, Y.; Yang, J.; Alzahrani, B.; Lv, G.; Alafif, T.; Barnawi, A.; Chen, M. Abnormal behavior learning based on edge computing toward a crowd monitoring system. *IEEE Netw.* **2022**, *36*, 90–96. [CrossRef]
35. Wang, K.; Liu, W.; He, X.; Qin, L.; Wu, X. A motion feature descriptor for abnormal behavior detection. *Comput. Sci.* **2020**, *47*, 119–124. [CrossRef]
36. Qin, Y.; Shi, Y. Human behavior recognition based on two-stream network fusion and spatio-temporal convolution. *Comput. Technol. Autom.* **2021**, *40*, 140–147. [CrossRef]
37. Jiang, J.; Wang, X.Y.; Gao, M.; Pan, J.; Zhao, C.; Wang, J. Abnormal behavior detection using streak flow acceleration. *Appl. Intell.* **2022**, *52*, 10632–10649. [CrossRef]
38. Clarke, K.C. Cellular automata and agent-based models. In *Handbook of Regional Science*; Springer Berlin Heidelberg: Berlin/Heidelberg, Germany, 2021; pp. 1751–1766.
39. Li, C.; Bi, C.; Li, Z. Crowd evacuation model based on improved PSO algorithm. *J. Syst. Simul.* **2020**, *32*, 1000–1008. [CrossRef]
40. Chang, D.; Cui, L.; Huang, Z. A cellular-automaton agent-hybrid model for emergency evacuation of people in public places. *IEEE Access* **2020**, *8*, 79541–79551. [CrossRef]
41. Chebi, H.; Dalila, A.; Mohamed, K. Dynamic detection of abnormalities in video analysis of crowd behavior with DBSCAN and neural networks. *Adv. Sci. Technol. Eng. Syst. J.* **2020**, *1*, 56–63. [CrossRef]
42. Tay, N.C.; Connie, T.; Ong, T.S.; Goh, K.O.M.; Teh, P.S. A robust abnormal behavior detection method using convolutional neural network. In Proceedings of the Computational Science and Technology: 5th ICCST 2018, Kota Kinabalu, Malaysia, 29–30 August 2018; Springer Singapore: Singapore, 2019; pp. 37–47.
43. Wang, Q.; Gao, J.; Lin, W.; Li, X. NWPU-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2141–2149. [CrossRef]
44. Koonce, B. ResNet 50. In *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*; Apress: New York, NY, USA, 2021; pp. 63–72.
45. Behera, N.K.S.; Sa, P.K.; Muhammad, K.; Bakshi, S. Large-Scale Person Re-Identification for Crowd Monitoring in Emergency. In *IEEE Transactions on Automation Science and Engineering*; IEEE: New York, NY, USA, 2023.
46. Wang, S.; Pu, Z.; Li, Q.; Wang, Y. Estimating crowd density with edge intelligence based on lightweight convolutional neural networks. *Expert Syst. Appl.* **2022**, *206*, 117823. [CrossRef]
47. Dai, F.; Huang, P.; Mo, Q.; Xu, X.; Bilal, M. Song, H. ST-InNet: Deep spatio-temporal inception networks for traffic flow prediction in smart cities. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 19782–19794. [CrossRef]
48. Liu, W.; Luo, W.; Li, Z.; Zhao, P.; Gao, S. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies. *IJCAI* **2019**, *3*, 3023–3030.
49. Bhuiyan, M.R.; Abdullah, J.; Hashim, N.; Farid, F.A.; Uddin, J. Hajj pilgrimage abnormal crowd movement monitoring using optical flow and FCNN. *J. Big Data* **2023**, *10*, 86. [CrossRef]
50. Garg, S.; Sharma, S.; Dhariwal, S.; Priya, W.D.; Singh, M.; Ramesh, S. Human crowd behaviour analysis based on video segmentation and classification using expectation–maximization with deep learning architectures. *Multimed. Tools Appl.* **2024**, 1–23. [CrossRef]
51. Singh, K.; Rajora, S.; Vishwakarma, D.K.; Tripathi, G.; Kumar, S.; Walia, G.S. Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. *Neurocomputing* **2020**, *371*, 188–198. [CrossRef]
52. 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.* **2019**, *172*, 88–97. [CrossRef]
53. Hu, X.; Dai, J.; Huang, Y.; Zhang, L.; Chen, W.; Yang, G.; Zhang, D. A weakly supervised framework for abnormal behavior detection and localization in crowded scenes. *Neurocomputing* **2020**, *383*, 270–281. [CrossRef]
54. Wang, H.; Zhou, M. Crowd abnormal behavior detection based on optical flow and trajectory. *J. Jilin Univ. (Eng. Ed.)* **2020**, *50*, 2229–2237. [CrossRef]
55. Hu, X.; Chen, Q.; Yang, L.; Yu, J.; Tong, X. Crowd abnormal behavior detection and location based on deep spatio-temporal convolutional neural network. *Appl. Res. Comput.* **2020**, *37*, 891–895. [CrossRef]
56. Gong, M.; Zeng, H.; Xie, Y.; Li, H.; Tang, Z. Local distinguishability aggrandizing network for human anomaly detection. *Neural Netw.* **2020**, *122*, 364–373. [CrossRef] [PubMed]
57. Pinaya, W.H.L.; Vieira, S.; Garcia-Dias, R.; Mechelli, A. Autoencoders. In *Machine Learning*; Academic Press: Cambridge, MA, USA, 2020; pp. 193–208.
58. Tyagi, B.; Nigam, S.; Singh, R. A review of deep learning techniques for crowd behavior analysis. *Arch. Comput. Methods Eng.* **2022**, *29*, 5427–5455. [CrossRef]
59. Xu, T.; Tian, C. Review of abnormal behavior detection of crowd based on deep learning. *Comput. Sci.* **2021**, *48*, 125–134. [CrossRef]
60. Zhang, X.; Ma, D.; Yu, H.; Huang, Y.; Howell, P.; Stevens, B. Scene perception guided crowd anomaly detection. *Neurocomputing* **2020**, *414*, 291–302. [CrossRef]
61. Zhou, S.; Shi, R.; Wang, L. Extracting macroscopic quantities in crowd behaviour with deep learning. *Phys. Scr.* **2024**, *99*, 065213. [CrossRef]

62. Nguyen, T.N.; Meunier, J. Anomaly detection in video sequence with appearance-motion correspondence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 1273–1283.
63. Wei, H.; Li, K.; Li, H.; Lyu, Y.; Hu, X. Detecting video anomaly with a stacked convolutional LSTM framework. In *International Conference on Computer Vision Systems*; Springer International Publishing: Cham, Switzerland, 2019; pp. 330–342.
64. Xiao, J.; Guo, H.; Xie, H.; Zhao, T.; Shen, M.; Wang, Y. Probabilistic memory autoencoder network for abnormal behavior detection in surveillance videos. *J. Softw.* **2023**, *34*, 4362–4377.
65. Nawaratne, R.; Alahakoon, D.; De Silva, D.; De Silva, D.; Yu, X. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Trans. Ind. Inform.* **2019**, *16*, 393–402. [CrossRef]
66. Roka, S.; Diwakar, M. Deep stacked denoising autoencoder for unsupervised anomaly detection in video surveillance. *J. Electron. Imaging* **2023**, *32*, 033015. [CrossRef]
67. Li, J.; Huang, Q.; Du, Y.; Zhen, X.; Chen, S.; Shao, L. Variational abnormal behavior detection with motion consistency. *IEEE Trans. Image Process.* **2021**, *31*, 275–286. [CrossRef]
68. Ashfahani, A.; Pratama, M.; Lughofer, E.; Ong, Y.S. DEVDAN: Deep evolving denoising autoencoder. *Neurocomputing* **2020**, *390*, 297–314. [CrossRef]
69. Wang, J.; Xia, L. Abnormal behavior detection in videos using deep learning. *Clust. Comput.* **2019**, *22* (Suppl. S4), 9229–9239. [CrossRef]
70. Wang, T.; Qiao, M.; Lin, Z.; Li, C.; Snoussi, H.; Liu, Z.; Choi, C. Generative neural networks for anomaly detection in crowded scenes. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 1390–1399. [CrossRef]
71. Lv, C.; Shen, F.; Zhang, Z.; Zhang, F. Review of the Research Status of Image Anomaly Detection. *Acta Autom. Sin.* **2022**, *48*, 1402–1428. [CrossRef]
72. Song, H.; Sun, C.; Wu, X.; Chen, M.; Jia, Y. Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. *IEEE Trans. Multimed.* **2019**, *22*, 2138–2148. [CrossRef]
73. Chen, D.; Yue, L.; Chang, X.; Xu, M.; Jia, T. NM-GAN: Noise-modulated generative adversarial network for video anomaly detection. *Pattern Recognit.* **2021**, *116*, 107969. [CrossRef]
74. Tang, Y.; Zhao, L.; Zhang, S.; Gong, C.; Li, G.; Yang, J. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognit. Lett.* **2020**, *129*, 123–130. [CrossRef]
75. Lee, S.; Kim, H.G.; Ro, Y.M. BMAN: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Trans. Image Process.* **2019**, *29*, 2395–2408. [CrossRef] [PubMed]
76. Schlegl, T.; Seeböck, P.; Waldstein, S.M.; Langs, G.; Schmidt-Erfurth, U. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med. Image Anal.* **2019**, *54*, 30–44. [CrossRef]
77. Meng, B.; Wang, L.; Li, D.W. Detection Method for Crowd Abnormal Behavior Based on Long Short-Term Memory Network. In Proceedings of the Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the 16th International Conference on IIHMSP in Conjunction with the 13th International Conference on FITAT, Ho Chi Minh City, Vietnam, 5–7 November 2020; Springer: Singapore, 2021; Volume 1, pp. 305–313.
78. Wu, G.; Guo, Z.; Li, L.; Wang, C. Video Anomaly Event Detection Fusing FCN and LSTM. *J. Shanghai Jiao Tong Univ.* **2021**, *55*, 607.
79. Sabih, M.; Vishwakarma, D.K. Crowd anomaly detection with LSTMs using optical features and domain knowledge for improved inferring. *Vis. Comput.* **2022**, *38*, 1719–1730. [CrossRef]
80. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**. [CrossRef]
81. Ye, Z.; Li, Y.; Cui, Z.; Liu, Y.; Li, L.; Wang, L.; Zhang, C. Unsupervised Video Anomaly Detection with Self-Attention Based Feature Aggregating. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; IEEE: New York, NY, USA, 2023; pp. 3551–3556.
82. Zhang, H.; Fang, X.; Zhuang, X. Video Human Abnormal Behavior Detection Model of Autoencoder Fusing Attention Mechanism. *Laser J.* **2023**, *44*, 69–75. [CrossRef]
83. Zhang, J.; Qi, X.; Ji, G. Self Attention Based Bi-Directional Long Short-Term Memory Auto Encoder for Video Anomaly Detection. In Proceedings of the 2021 Ninth International Conference on Advanced Cloud and Big Data (CBD), Xi'an, China, 26–27 March 2022; IEEE: New York, NY, USA, 2022; pp. 107–112.
84. Zhang, W.; Wang, G.; Huang, M.; Wang, H.; Wen, S. Generative Adversarial Networks for Abnormal Event Detection in Videos Based on Self-Attention Mechanism. *IEEE Access* **2021**, *9*, 124847–124860. [CrossRef]
85. Singh, R.; Sethi, A.; Saini, K.; Saurav, S.; Tiwari, A.; Singh, S. Attention-Guided Generator with Dual Discriminator GAN for Real-Time Video Anomaly Detection. *Eng. Appl. Artif. Intell.* **2024**, *131*, 107830. [CrossRef]
86. Sharma, V.K.; Mir, R.N.; Singh, C. Scale-Aware CNN for Crowd Density Estimation and Crowd Behavior Analysis. *Comput. Electr. Eng.* **2023**, *106*, 108569. [CrossRef]
87. Zhao, R.; Wang, Y.; Jia, P.; Zhu, W.; Li, C.; Ma, Y.; Li, M. Abnormal Behavior Detection Based on Dynamic Pedestrian Centroid Model: Case Study on U-Turn and Fall-Down. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 8066–8078. [CrossRef]
88. UCSD Anomaly Detection Dataset. Available online: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm (accessed on 20 September 2024).
89. UMN Crowd Dataset. Available online: https://mha.cs.umn.edu/proj_events.shtml#crowd (accessed on 23 September 2024).

90. Cao, C.; Lu, Y.; Zhang, Y. Context Recovery and Knowledge Retrieval: A Novel Two-Stream Framework for Video Anomaly Detection. *IEEE Trans. Image Process.* **2024**, *33*, 1810–1825. [CrossRef]
91. Wang, L.; Wang, X.; Li, M.; Liu, F. Anomaly Detection Method Based on Temporal Spatial Information Enhancement. *Meas. Sci. Technol.* **2023**, *35*, 035410. [CrossRef]

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