Applied Intelligence (2025) 55:904
https://doi.org/10.1007/s10489-025-06797-4

Deep learning applied for abnormal human behavior recognition in video surveillance systems: A systematic review

Olfa Saket[1] · Anis Ben Aicha[2] · Habib Fathallah[1]

Accepted: 12 July 2025

Abstract
The automatic recognition of human behavior is attracting more research attention, especially with the rapid progress of neural networks in recent years. The efficient detection and recognition of abnormal human behavior (AHB) are critical components of intelligent video surveillance systems, as they ensure human security and create safe environments. This paper presents a systematic literature review (SLR) of 140 studies published between 2016 and 2024, focusing on the application of deep learning for AHB recognition in videos. We formulate eight key research questions (RQs) that explore the types of AHB addressed in the literature, the historical progression of deep learning models, pre-trained architectures, existing methods, real-time applications, datasets, performance metrics, and future research directions. This review paper serves as a valuable guide for both academia and industry professionals seeking to understand this field and explore the top emerging trends.

Keywords Deep learning · Human behavior · Anomaly recognition · Intelligent video surveillance

1 Introduction

In the modern world, intelligent video surveillance systems have become a critical component, driven by the increasing demand for automated visual data analysis across various domains. One of the most important tasks provided by these systems is the recognition of human behavior. The ability to automatically detect and recognize human behavior has become an important aspect of intelligent video systems, as it has various practical applications in the real world such as behavior analysis and understanding [1], action recognition [2], and behavior surveillance [3]. Indeed, monitoring human behavior in video surveillance systems is essential for enhancing public security and creating safer environments. The recognition of AHB involves identifying unusual actions and events that may lead to threats and emergencies in various contexts. Examples include fight detection [4], crime identification [5], fall detection [6], and driver behavior recognition [7].

Many researches have been conducted in this field over the years. Traditional approaches typically rely on handcrafted feature extraction techniques, such as Histogram of Oriented Gradients (HOG) [8], Scale-Invariant Feature Transform (SIFT) [9], Space-Time Interest Points (STIP) [10], and histograms of optical flow orientation (HOFO) [11]. These features are then processed by traditional machine learning algorithms, including Support Vector Machines (SVM) [12], Random Forests [13], and k-Nearest Neighbors (kNN) [14]. These approaches are limited by their linear nature and inability to handle large volumes of unstructured data such as videos. The rise of deep learning has revolutionized this field, with their capability to automatically extract spatial, temporal, and contextual features, and learn complex patterns from large amount of data. Deep learning models have shown superior performance

Anis Ben Aicha and Habib Fathallah contributed equally to this work.

✉ Olfa Saket
olfa.saket@fsb.ucar.tn

Anis Ben Aicha
anis.benaicha@supcom.tn

Habib Fathallah
habib.fathallah@fsb.ucar.tn

[1] University of Carthage, FSB, LR21ES23, IDEA, Bizerte, Tunisia

[2] University of Carthage, SUPCOM, LR11TIC01, COSIM, Ariana, Tunisia

Published online: 02 August 2025

compared to traditional methods, demonstrating their effectiveness in handling complex visual tasks including abnormal behavior recognition.

This paper provides a SLR of studies that detect and recognize AHB through deep learning techniques, focusing on research published between 2016 and 2024. The main contribution of this study is an in-depth analysis of the various types of AHB addressed in the literature, with particular focus on those that require further research attention. Another significant contribution consists of a comprehensive examination of deep learning models used, including recent and advanced architectures. Additionally, this work outlines the full steps of a video-based AHB recognition system, from data preprocessing to feature extraction, modeling, and evaluation. The study also describes the datasets commonly used in the literature, evaluation metrics employed, performance obtained, and identifies several key future research directions. These contributions are summarized as follows:

1. Categorization of AHB to review those that have already been addressed and identify human behaviors that require more attention.
2. An overview of deep learning models applied for AHB detection and recognition in video surveillance.
3. An in-depth examination of existing approaches along with a short overview of techniques and paradigms employed for real-time detection.
4. A thorough study of the commonly used datasets, evaluation metrics and performance obtained.
5. Identification of research gaps and future directions in video-based AHB recognition systems.

The remainder of the paper is organized as follows: Section 2 provides a literature review of related previous works. Section 3 describes the research methodology used in this paper. Section 4 presents the results of this study. Section 5 provides a comparison of our work with the existing SLR. Finally, Section 6 concludes the paper with a summary of the findings.

2 Related reviews

Abnormal behavior recognition is one of the most important elements in video-based behavior analysis. The increasing number of surveillance cameras, along with advancements in technology for automatically processing visual data during the last years, encourages researchers to more contribute to the development of solutions for identifying security threats and recognizing criminal activities. Many reviews have been conducted to summarize the existing studies, focusing on various techniques such as statistical methods, machine learning and deep learning. These reviews can be divided into two categories: Traditional Literature Reviews (TLRs) and Systematic Literature Reviews (SLRs). A TLR provides a comprehensive overview of existing studies by analyzing and synthesizing their results, while a SLR follows a predefined methodology to systematically identify, analyze and synthesize relevant works in order to answer specific research questions [15]. In fact, there are significantly more TLRs compared to SLRs. In the following two subsections, we provide a brief summary of the recent traditional and systematic reviews in the field.

2.1 Traditional literature reviews

Many TLRs have been published over the past few years. In 2021, Nayak et al. [16] published an extensive study on deep learning-based methods for video anomaly detection and localization. The authors offered a comparative analysis of the existing methods in terms of datasets, computational infrastructure, and performance metrics. They also emphasized several research challenges associated with datasets, computational complexity, evaluation methodologies, and environmental challenges. In 2022, Patrikar and Parate [17] explored existing anomaly detection techniques that have been applied for intelligent video surveillance systems and discussed the adoption of edge computing for real-time applications. Moreover, several literature reviews were published in 2023. For example, Berroukhan et al. [18] presented a comparative analysis of deep learning-based methods for video anomaly detection. This analysis included reconstruction error, future frame prediction, classifiers, and scoring methods. Duong et al. [19] published a literature revue that provides a detailed discussion on pre-prossessing techniques, feature engineering methods and deep learning-based approaches utilized for detecting AHB. In [20], Huang et al. discussed various deep learning models used for real-time anomaly detection in internet of things video surveillance systems. Several studies have also been published in 2024. Liu et al. [21] proposed an hierarchical taxonomy to systematically categorize the existing deep generalized video anomaly event detection models based on their supervised methods, inputs of the model, and the network structure. Wastupranata et al. [22] provided a comprehensive overview of deep learning techniques for detecting AHB in surveillance video streams. The study presented popular used datasets, classified existing techniques into unsupervised, partially supervised, and fully supervised approaches, and highlighted open research challenges in the field. Negre et al. [23] proposed a thorough systematic mapping study on deep learning-based violence detection in videos. The

authors categorized the existing approaches based on detection algorithms, specific challenges, datasets, key-frame extraction methods, and model performance.

2.2 Systematic literature reviews

On the other hand, a few works have been systematically reviewed the existing state-of-the-arts researches. In 2022, Omarov et al. [24] presented a systematic review on violence detection methods for video surveillance applications. They reviewed 154 studies published between 2012 and 2021. The SLR involved an in-depth examination of traditional machine learning, SVM, and deep learning-based approaches as well as video features and descriptors, datasets, evaluation metrics, and challenges. In 2023, Bouhsissin et al. [25] provided a comprehensive systematic review for driver behavior classification systems. Their analysis covered 93 papers published from 2015 to 2022. The paper classified driver behaviors into five categories and provided further details about prepossessing techniques, features selection and description, banchmark datasets, traditional and deep learning techniques and performance metrics utilized during the reviewed period. In 2024, Samaila et al. [26] conducted a SLR for video-based abnormal behavior detection systems by analyzing 530 studies published between 2003 and 2023. The review provided a comprehensive analysis of classical, traditional, and deep learning-based approaches. It also highlighted key research challenges and outlined several future research directions, including real-time anomaly detection, online learning and adaptation and Multi-camera anomaly detection.In the same year, Gaya-Morey et al. [27] proposed a SLR that addressed both fall detection and human activity recognition for the elderly using visual data. The authors analyzed 151 research studies published between 2019 and 2023. The study discussed various aspects including data types, deep learning architectures, hardware used, and privacy protection strategies

Although these existing works provided systematic and comprehensive overviews of the research published in the field, several significant limitations are identified. Most of these studies focus on specific types of anomalies such as violence [24], driver behavior [25], and falls[27] rather than encompassing the full range of abnormal behaviors. In addition, these reviews often overlook recent advances in deep learning, particularly the use of Transformer-based models. Therefore, there is a need for a complete and up-to-date systematic review that covers a wide rang of AHB and incorporates the latest developments in deep learning. This work stands out as an extensive review, bridging the research gaps identified in previous studies by systematically reviewing recent advancements in video-based AHB recognition research.

3 Research methodology

The study is conducted following the methodology proposed by Kitchenham and Charters [28]. To guide the review process and minimize the likelihood of research bias, we developed a review protocol composed of five main stages: RQs, search strategy, study selection, data extraction, and synthesis strategy. The details of each stage are presented in the following subsections.

3.1 Research questions

Table 1 lists the RQs addressed throughout this SLR. We formulate eight RQs in order to systematically analyze and synthesize research studies published between 2016 and 2024. RQ1 categories the AHB that have been addressed in the identified research studies. RQ2 explores the deep learning models utilized during the selected period. RQ3 provides an overview of the pre-trained models used. RQ4 investigates the key steps of a video-based AHB recognition system and identifies which methods are most commonly employed at each step. RQ5 discusses the computing paradigms and techniques used to enable real-time detection. RQ6 demonstrates the commonly used datasets. RQ7 analyzes the performance metrics utilized to evaluate models and their obtained results. RQ8 identifies some future research directions to overcome the current limitations in the field.

Table 1 Research questions formulated for this SLR

ID	Research question
RQ1	What kinds of AHB are addressed in the identified research studies?
RQ2	What is the historical progression of deep learning models used for AHB recognition in videos?
RQ3	What are the most commonly used pre-trained deep learning models?
RQ4	What are the key steps of a video-based AHB recognition system, and which methods are most commonly employed at each step?
RQ5	What computing paradigms and techniques are used to enable real-time AHB detection?
RQ6	What are the commonly used datasets to train and evaluate models?
RQ7	What are the performance metrics considered for model evaluation?
RQ8	What future research directions can address the current limitations of AHB recognition systems and enhance their performance in real-world environments?

3.2 Search strategy

3.2.1 Queries

To prepare our search queries, we started by identifying three main keywords related to the formulated RQs. These keywords are abnormal behavior, deep learning and video. Then, we generated a set of derivatives for each one. Based on these derivative sets, we constructed four search queries. Each query is composed of three sub-queries. For each derivative set, we selected multiple terms and connected them with the OR operator, formulating a sub-query. Finally, we combined the three sub-queries obtained with the AND operator to formulate a complete search query. Table 2 lists the four search queries we prepared to collect relevant papers from academic databases. The first query incorporates all main keywords and their derivatives, while the remaining three queries are limited to a maximum of eight operators

3.2.2 Database selection

Seven online academic databases were consulted as source for this SLR: IEEE Xplore, ACM Digital Library, Google Scholar, ScienceDirect, Springer Link, ResearchGate, and Wiley. All formulated queries were executed in the mentioned databases, except for the first query, which was executed only in the ACM Digital Library, IEEE Xplore, and Wiley due to the limited number of operators allowed in the other databases.

Table 2 Search Queries

ID	Search query
Q1	(“abnormal behavior" OR “human behavior" OR “anomaly detection" OR “behavior recognition" OR “behavior detection" OR “aberrant behavior" OR “suspicious behavior" OR “behavior surveillance" OR “behavior monitoring") AND (“deep learning" OR “neural network" OR “deep network" OR “Deep architecture" OR “automatic detection" OR “automatic recognition") AND (“video" OR “video surveillance" OR “video sequence" OR “video system" OR “CCTV" OR “video monitoring" OR “camera" OR “surveillance camera" OR “video analysis")
Q2	(“anomaly detection" OR “behavior recognition" OR “human behavior") AND (“deep learning" OR “neural network" OR “automatic detection") AND (“video surveillance" OR “video sequence" OR “video system")
Q3	(“suspicious behavior" OR “unusual behavior" OR “behavior recognition") AND (“Depp learning" OR “Deep architecture" OR “automatic recognition") AND (“video monitoring" OR “CCTV" OR “surveillance camera")
Q4	(“abnormal behavior" OR ‘human behavior’) AND (“deep learning" OR “neural network") AND (“video" OR “camera")

3.3 Study selection

A total of 1032 studies, including 969 journal articles and 163 conference papers, were collected based on the search queries mentioned above. These studies were filtered by applying inclusion and exclusion criteria to choose only the most pertinent ones for this work. The screening processes of the collected papers is detailed in Table 3 and Fig. 1. The inclusion and exclusion criteria applied are as follows:

Inclusion criteria IC1: Paper published from 2016 to 2024.
IC2: Paper addresses the use of deep learning for the detection and recognition of AHB in videos.
IC3: The predefined keywords exist at least in the title, keywords, abstract or conclusion of the paper.

Exclusion criteria EC1: Paper is not written in English language.
EC2: Paper is not a journal or conference paper.
EC3: Paper published in a journal not indexed by Scimago Journal, Country Rank (ScimagoJR), Journal Citation Reports (JCR) or CORE Rankings Portal.
EC4: The journal rank is greater than Q3.
EC5: Conference paper published in a conference not indexed by CORE Rankings Portal.
EC6: The conference rank is not in A+, A, B.

4 Results

The present section provides a comprehensive systematic analysis of the identified research studies. A total of 1032 research papers that addressed AHB detection in videos using deep learning techniques were collected. After the application of the selection criteria, only 140 studies were chosen, including 118 journal articles and 22 conference papers. These studies were released between 2016 and 2024. Figure 2 shows the annual distribution of the selected papers. The next subsections present all the answers to the eight RQs.

Table 3 Applying inclusion and exclusion criteria process

Step	Applied criteria	Initial papers	Excluded papers	Remaining papers
S1	IC1, EC1, EC2	1038	236	802
S2	IC2, IC3	802	186	616
S3	Duplicates	616	174	442
S4	EC3, EC4, EC5, EC6	442	278	164
S5	Full text, IC2, IC3	164	24	140

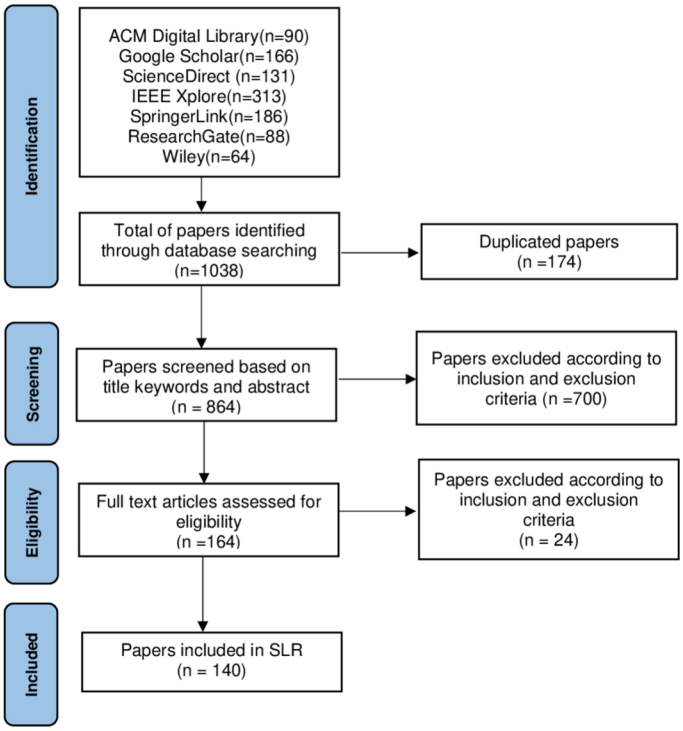

Fig. 1 PRISMA flow diagram

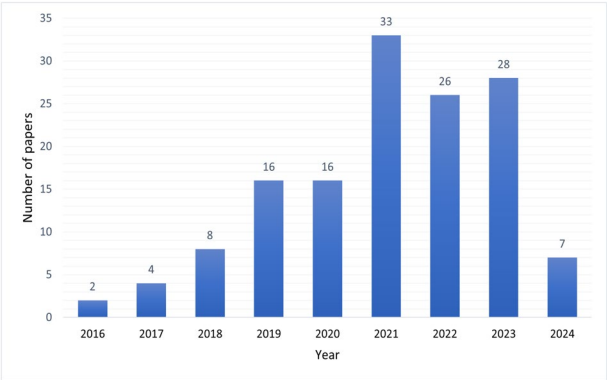

Fig. 2 Annual distribution of selected papers

4.1 What kinds of AHB are addressed in the identified research studies?(RQ1)

In fact, any action or event that could have a negative impact on human safety and security can be considered abnormal behavior. To gain a comprehensive understanding of the types of abnormal behavior found in the reviewed studies, we organized them into four categories: aggressive behavior, suspicious behavior, driver behavior, and health-related behavior, as shown in Fig. 3. It is important to note that this classification methodology reflects our own perspective in exploring abnormal behavior. The goal is to provide a structured and coherent view of the behavior already addressed in the literature.

4.1.1 Aggressive behavior

Aggressive behavior includes actions or events made by humans that cause harm or damage to others. For this type of anomaly, there is a real danger occurring. Multiple studies, including Mohtavipour et al. [29], Ullah et al. [30], Hussain et al. [31], Asad et al. [32], Serrano et al. [33], Asad et al. [34], Samuel et al. [35], Irfanullah et al. [36], and Magdy et al. [37], proposed deep learning-based methods for fight detection. Garcia-Cobo and SanMiguel [38] introduced a deep learning approach to detect violent crimes in public places. Hlavatá et al. [39] developed an approach for recognizing abnormal human activities, including begging, drunkenness, fighting, harassment, hijacking, knife-related threats, pollution, property damage, robbery, and terrorism. In [40], Sultani et al. proposed a weakly supervised approach to detect and recognize real-world aggressive activities such as abuse, arrest, arson, assault, burglary, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism. Alia et al. [41] developed a framework for early detection of pushing in crowded event entrances. Buttar et al. [42] proposed hybrid deep learning models to detect and recognize six abnormal human activities that are running, punching, falling, snatching, kicking, and shooting. Several studies, including Huszár et al. [43], Khan et al. [44], and Sudhakaran and Lenz [45], developed approaches for detecting various violent behaviors such as crimes, shootings, fighting, kicking, beating, stabbing, and

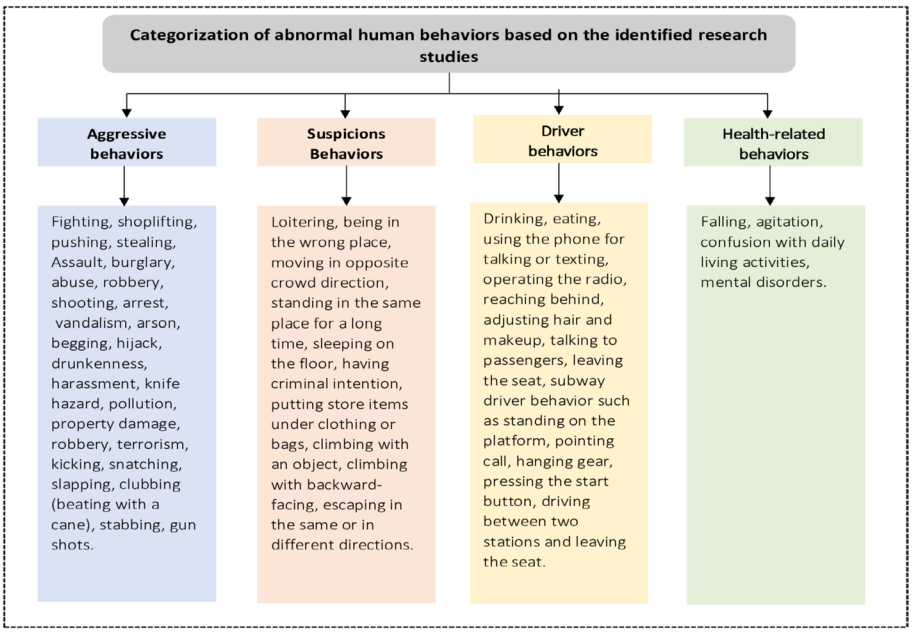

Fig. 3 Categorization of abnormal human behaviors

others. Bala and Kaushal [46] proposed a deep learning model for jaywalking detection and localization.

4.1.2 Suspicious behavior

Suspicious behavior refers to behavior that draws the attention of surveillance staff and concerns them. This type of behavior is considered annoying or inappropriate but does not involve aggression. Several studies, such as Martínez-Mascorro et al. [47] and Ansari and Singh [5], introduced approaches to identify criminal intention that may lead to shoplifting activities. Alafif et al. [48] proposed a behavior detection method to detect suspicious behaviors from the annual hajj event including sleeping on the floor, standing in the same place, and crossing or moving in different crowd direction. Direkoglu [49] focused on detecting global abnormal crowd events including sudden escape of people in the same or different directions. Mehmood [50] proposed a method to detect loitering in uncrowded scenes.

4.1.3 Driver behavior

Any driving behavior that represents a risk to the vehicle, its occupants, passengers, other drivers, or roadside infrastructure can be classified as abnormal. Celaya-Padilla et al. [51] proposed an automated method to detect distracted drivers using their mobile phones. Huang et al. [52] introduced three deep learning-based fusion models to detect abnormal driver behaviors, including texting, talking on the phone, operating the radio, drinking, reaching behind, adjusting hair and makeup, and talking to passengers. Huang et al. [53] developed a recognition model that automatically recognize behaviors of subway drivers such as pressing the start button, leaving the seat and driving between two stations. Zhang et al. [54] proposed a driver behavior recognition system that identify different types of driver behavior, including texting, eating, talking, using a mobile phone, drinking, and preparing to drive and finishing driving.

4.1.4 Health-related behavior

Health-related behavior refers to strange behavior performed by individuals or patients that may indicate physical or mental health issues. We find that several studies including Chhetri et al. [55], Rajavel et al. [56], Fei et al. [6], and Li et al. [57] proposed fall detection methods for the elderly. khan et al [58] developed a model to detect agitation and recognize individuals with dementia. Hao et al. [59] introduced an approach that recognize and localize abnormal behaviors for individuals with mental disorders.

4.2 What is the historical progression of deep learning models used for AHB recognition in videos? (RQ2)

In this section, we explore and examine the evolution of deep learning models used for detecting and recognizing abnormal behaviors in video sequences. We found that 86 % of the identified studies utilized exclusively deep learning models for anomaly detection, while 14% combined deep learning with traditional methods. As depicted in Fig. 4, Convolutional Neural Networks (CNNs) and Autoencoders (AEs) have been consistently employed in the field since 2016. Recurrent Neural Networks (RNNs), encompassing

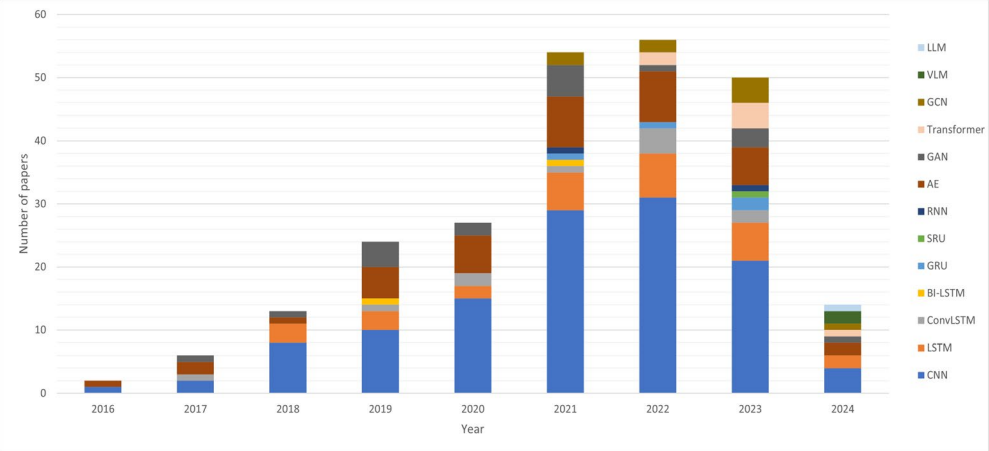

Fig. 4 Distribution of deep learning Models Used in reviewed studies

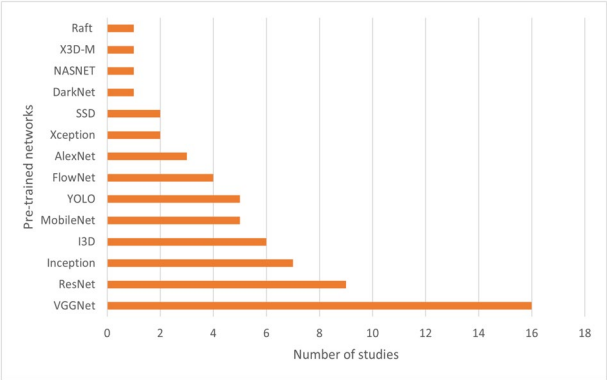

Fig. 5 Distribution of the pre-trained networks

variants such as Long Short-Term Memory (LSTM), Convolutional LSTM (Conv-LSTM), Bidirectional LSTM (Bi-LSTM), Gated Recurrent Unit (GRU) and Simple Recurrent Unit (SRU), emerged as popular choices from 2017 onwards. Generative Adversarial Networks (GANs) were introduced in 2017. Graph Convolutional Networks (GCNs) have been utilized in this domain since 2021, while Transformers [60] have gained traction, with their application commencing in 2022. More recently, Vision-Language Models (VLMs) have been used since 2024, followed by the integration of Large Language Models(LLMs) in the same year.

4.3 What are the most commonly used pre-trained deep learning models? (RQ3)

Recently, transfer learning has received significant attention, particularly in fields with limited high-quality training data [61, 62]. It involves transferring a model's pre-trained knowledge, learned from general tasks, to new specific tasks. A preexisting model trained on a large-scale dataset can be directly used to perform a specific task without any fine-tuning process. This transfer learning approach typically reduces training time and computational requirements, but limits the model's ability to identify complex, task-specific patterns. Another way of appling transfer learning is fine tuning. The fine-tuning adapts pre-trained models for specific tasks by updating their parameters using smaller, domain-specific datasets. This approach focuses on improving task-specific performance while using less data. Figure 5 demonstrates the pre-trained deep learning models used in the reviewed studies, along with their frequency of usage. The graphic shows that VGGNet [63] is the most frequently used network due to its ability to extract robust visual features. Following VGGNet, ResNet [64] is also widely employed because of its residual mechanisms. Next, Inception [65] is utilized for its capability of extracting information at different scales through its convolutional layers with various filter sizes. Then, I3D [66] is chosen for their ability to handle temporal information, while MobileNet [67] and YOLO [68], are utilized for their computational efficiency in real-time object detection. In addition, FlowNet [69] is commonly used for deep optical flow estimation. Some networks are still limited in their use, including AlexNet [70], Xception [71], SSD [72], DarkNet, NASNET, X3D-M [73], and Raft [74]. It is important to note that ImageNet [75] is the most commonly used dataset for pre-training these networks due to its large scale and diverse set of labeled images.

4.4 What are the key steps of a video-based AHB recognition system, and which methods are most commonly employed at each step? (RQ4)

As illustrated in Fig. 6, the automatic recognition of AHB in videos is generally performed through six fundamental steps. The first step is the model input. The second step includes the data preprocessing. The third step involves human detection and tracking. The fourth step comprises feature extraction

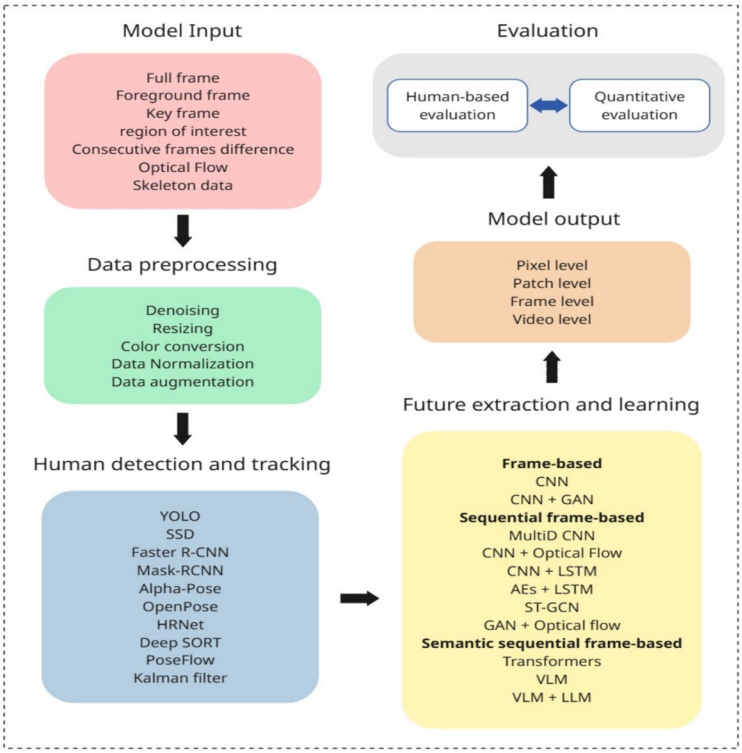

Fig. 6 Pipeline of a video based AHB recognition system

and learning. The fifth step is the output. Finally, the sixth step consists of the evaluation process. The following subsections provide a brief explanation of these steps and highlight the most commonly used methods for each one.

4.4.1 Model input

The choice of input data for detecting anomalies is crucial, as it directly affects the accuracy and effectiveness of the model. Depending on the approach and detection objectives, inputs can range from a full video frame to skeleton data. In this subsection, we identify and analyze the different types of input used in the reviewed studies.

- **Full frame** Full frame involves using the entire video frame as input to the model. By analyzing the full frame, the model can detect anomalies within the entire visual context of the frame. This technique ensures that all visual information is processed, but it may requires substantial computational resources, especially when dealing with high-resolution frames or large amount of video data. [76–82].
- **Foreground frame** Foreground frame involves removing the background from the frame and using only the foreground as input. By removing the background, the model focuses on the moving entities, which improves the accuracy and reduces the computational and storage requirements. [83].
- **key frames** Key frames are the most commonly used input for video-based abnormal behavior recognition. Key frames selection techniques consist of reducing the number of frames to be processed by selecting only those that represent significant changes in the input video. The use of key frames as input reduces less informative and redundant frames from the video, which helps to minimize the computational and storage costs and enhances the model performance [84–86].
- **Region of interest** Another input frequently used in the reviewed studies is the region of interest. Instead of using the full video frame, active regions can be extracted and used as input to the model. By focusing only on specific areas, which may include regions around objects or individuals of interest, the model can improve its accuracy. This technique allows computational resources to be concentrated on the relevant parts of the frame [87, 88].
- **Frame patches** Frame patches consist of dividing a video frame into smaller pieces of fixed size called patches. This approach allows the model to process local regions, which can significantly enhance its ability to capture fine-grained details and detect localized anomalies[89–95].

- **Consecutive frames difference** Consecutive frames difference refers to the process of detecting intensity changes between two or more consecutive frames. This is achieved by subtracting the pixel values of one frame from those of the following frame. This operation can reduce the volume of data processed during training by focusing only on the significant changes within a video sequence.[36, 45, 48, 96].
- **Optical Flow** Many research studies used motion information obtained through optical flow as input to the model. Optical flow represents the speed and direction of moving entities by calculating the displacement of pixels between consecutive frames. This technique allows the model to focus on the moving entities rather than static background elements. Various traditional [55, 97–104] and deep learning-based methods [105–107] are used in the literature to estimate optical flow.
- **Skeleton data** Another commonly used input is the human skeleton, which is extracted using pose estimation methods. The aim of these methods is to identify key points on the human body and represent them as coordinates in either 2D or 3D dimensions. This representation allows the model to learn the spatial relationships between different body parts, which is important for detecting human movements [108–113].

4.4.2 Data preprocessing

The quality of the training data is one of the most critical factors that directly influences the performance of deep learning models. Data preprocessing refers to a set of techniques applied to prepare and transform raw data into a format suitable for training deep learning models. Five data preprocessing techniques have been identified from the selected studies including denoising, resizing, color conversion, normalization, and data augmentation. Resizing is the most commonly applied technique, followed by normalization, data augmentation, color conversion, and denoising. A detailed analysis of these techniques is provided in the following subsections.

- **Denoising** Denoising refers to the process of removing unwanted small objects from video frames to improve their quality. Several studies use Gaussian filtering which consists of applying a Gaussian filter to video frames to reduce noise and produce smoothed frames [96, 114, 115]. Other studies used bilateral filtering for edge-preserving smoothing [31, 36, 115–117].
- **Resizing** Resizing video frames is widely used as a preprocessing technique to standardize the dimensions of frames by removing borders that generally do not contain relevant details. This can reduce computational costs and processing time, while allowing the model to focus on the most relevant details of the frame. [39, 118–125].
- **Color conversion** Several researchers choose to reduce the number of video frame channels by converting frames from RGB to gray-scale. This approach can minimize computational costs and accelerate the training phase. Other researchers prefer to retain color information, which can be useful for detecting anomalies. Certain deep learning models impose a specific channel order for video frames, which need the swapping of color channels to match the model's input format [50, 114, 126–130].
- **Data Normalization** Data normalization is a preprocessing technique used to standardize the scale and distribution of data values. It ensures that the values are on a consistent scale, which helps the model converge more effectively during training. For video data, researchers generally normalize pixel values by scaling them to ranges such as 0 to 1 or -1 to 1. Another technique used is histogram normalization, which involves adjusting the histogram of pixel intensities to produce a more uniform distribution. [131–135].
- **Data augmentation** Data augmentation is widely employed, especially when using advanced models with high data requirements. It involves applying a set of transformations to the initial data to generate new instances. The transformation techniques used in the identified studies are random flipping, cropping, pixel dropout, noise additions, rotation, translation, zoom, size variation, color intensity adjustment, stride-based temporal augmentation, sliding windows, and coordinate transformation. These transformations allow the model to process different variants of the same data, which can enhance its performance [29, 125, 136–140].

4.4.3 Object detection and tracking

Object detection and tracking are essential components in video anomaly detection. Among object detection models selected from the identified studies, YOLO is the most widely used due to its speed and real-time efficiency. SSD is also frequently employed, as it offers a good balance between speed and accuracy. Faster R-CNN [141] and Mask-RCNN [142] are less used. For human pose extraction, models such as Alpha-Pose [143] and OpenPose [144] are frequently used, while HRNet [145] is less used. For object tracking, Deep SORT [146] is often employed due to its ability to efficiently track multiple objects in complex scenarios, followed by PoseFlow [147] for tracking human joints. Finally, the Kalman filter [148] is a traditional method that is still widely used.

4.4.4 Learning methods

To provide a structured and comprehensive overview of the learning methods proposed in the identified studies, we classify them into three main categories based on the type of information they leverage during the learning process. These categories are as follows: Frame-based methods, sequential frame-based methods, and semantic sequential frame-based methods. Frame-based methods identify static anomalies using spatial features extracted from individual frames, without incorporating any temporal information. These methods focus on identifying local relationships between pixels within a single frame. They detect anomalies based on localized visual features, but cannot capture the global relationships between pixels across the entire frame. Consequently, their performance is generally limited to detecting static anomalies that appear in a single frame. Sequential frame-based methods model temporal dependencies across sequences of frames to detect dynamic anomalies that unfold over time. These methods extend the capabilities of frame-based approaches by incorporating temporal tracking across consecutive frames. This additional temporal information significantly enhances their ability to detect dynamic anomalies. Semantic sequential frame-based methods learn context-aware representations of video frames, enabling semantic detection of anomalies. These methods are typically based on advanced architectures such as transformers. These architectures capture spatial relationships between different patches within each frame and model temporal dependencies across sequences of frames enabling semantic anomaly detection. Table 4 provides strengths and limitations of each category.

- **Frame-based methods**
  - **CNN-based methods** In 2018, Sabokrou et al. [92] combined a pretrained fully convolutional neural networks(FCNs) with Gaussian classifier for global anomaly detection. In 2020, Pang et al. [149] proposed an end-to-end approach for unsupervised video anomaly detection based on self-trained ordinal regression. Initially, pseudo-anomalous and normal frame sets were generated, and abnormal scores were computed using a pre-trained ResNet-50 followed by FCLs. Then, a self-training mechanism was employed to optimize the anomaly detection model. In 2022, Lalit et al. [76] used two convolutional layers for spatial feature extraction, followed by two FCLs where the first uses a Rectified Linear Unit and the second applies a sigmoid function for abnormal behavior detection in crowded scene. In the same year Irfanullah et al. [36] used a pre-trained CNN, MobileNets, followed by FCLs for real-time violence detection.
  - **CNN + GAN-based methods** In 2021, Chen et al. [150] designed a GAN-like architecture for unsupervised anomaly detection, composed of three modules. First, an image-to-image encoder-decoder network based on CNN and U-Net [151] was used for frame reconstruction. Then a CNN-based discrimination network was employed to select anomaly frames based on the reconstruction error maps. Finally, the output of the discrimination network was fed into an anomaly scoring estimation model to compute the anomaly score of the input frame.

- **Sequential frame-based methods**
  - **MultiD CNN-based methods** In 2018, Sultani et al. [40] introduced a deep multi-instance learning framework that predicts anomaly scores for video segments. First, the authors split each video into multiple segments. Then, a pre-trained 3D CNN [152] was employed to extract C3D features from the video segments. Finally, a fully connected network with a ranking loss was trained to assign anomaly scores to the segments. In 2019, Ullah et al. [153] introduced an approach to predict violent behavior in public spaces. First, a lightweight CNN was used for person detection. The frame sequences with the detected persons were then fed into a 3D CNN, which captured the spatio-temporal characteristics and passed them to a softmax classifier for the final predictions. In 2021, Li et al.[154] employed an AE to reconstruct individual frames and generate latent codes that capture appearance features. These latent codes, extracted from sequences of frames, were then stacked and fed into a 3D CNN, which was trained to predict the latent codes of future

Table 4 Learning methods: Strengths and limitations

Category	Strengths	Limitations
Frame-based methods	Simple, good performance on static anomalies, faster training time, lower computational cost, suitable for local devices	Ignore temporal dependencies
Sequential frame-based methods	Model short-term temporal dependencies, dynamic anomaly detection	Struggle with long events, lack semantic encoding, challenges in handling previously unseen anomalies
Semantic sequential frame-based methods	Model long-term dependencies through attention mechanisms, handle previously unseen anomalies, semantic representations of data, better performance	Higher computational and storage costs, unsuitable for local devices

frames. In 2022, Hao et al. [155] implemented a spatio-temporal consistency-enhanced network that uses a 3D-CNN encoder and a 2D-CNN decoder to capture feature representations and generate the next frame. Then, a 3D CNN-based discriminator was employed to measure the spatiotemporal consistency between the generated frame and its former frames. In 2023, Huszár et al. [43] introduced two models: a fine-tuned X3D-M model and a transfer learned X3D-M model. The first model fine-tuned the X3D-M architecture pre-trained on the Kinetics-400 dataset, while the second model used the pre-trained X3D-M to extract spatio-temporal features, which were then passed through FCLs and a sigmoid function for violence detection.

- **CNN + Optical Flow-based methods** In 2016, Zhou et al. [156] used optical flow to select spatio-temporal volumes of interest from video frames, which were then fed into a spatio-temporal CNN to extract spatio-temporal features and localize abnormal activities in crowded scenes. In 2020, Doshi and Yilmaz [157] introduced an online anomaly detection method using transfer learning and continual learning. YOLOv3 was employed to extract location and appearance features, while FlowNet2 captured motion information. The extracted features were then fed into a statistical k-nearest neighbor decision algorithm, which makes online decisions and continually updates its decision rule. In 2022, Mohtavipour et al. [29] proposed a violence detection framework that incorporates handcrafted features, including speed of movement, appearance, and representative images. These features were processed through a three-stream CNN: a spatial stream for learning environmental patterns, a temporal stream utilizing a modified optical flow to capture aggressive motion behaviors, and a spatiotemporal stream using motion energy images to extract discriminative features.
- **CNN + LSTM-based methods** In 2017, Sudhakaran and Lanz [45] used video frame differences as input to AlexNet for spatial features extraction, which were then processed by a Conv-LSTM to capture spatio-temporal features. The resulting features were passed through a series of FCLs for violence recognition. In 2018, Ding et al. [136] used Inception-v3 and LSTM to extract spatial and temporal information, followed by a Softmax classifier to identify unsafe behaviors at construction sites. In 2019, Asad et al. [32] employed a pre-trained VGG-16 to extract low-level and high-level feature maps from two consecutive frames. The extracted feature maps were fused and passed through residual blocks to learn enriched spatial representations. Then, an LSTM was employed to model global temporal dependencies followed by FCLs to detect violent behaviors. In 2020, Feng et al. [158] developed a spatiotemporal fall detection method that identify spatial and temporal locations of fall events in complex public scenes. The authors employed YOLOv3 for pedestrian detection and used Deep SORT for tracking. Then, a VGG16 was applied to extract features from each trajectory, which were then fed into an attention-guided LSTM model for final predictions. In 2021, Ullah et al. [159] used a pre-trained ResNet-50 to extract features from consecutive video frames. These features were then fed into a BD-LSTM classifier for anomaly classification. In 2022, Vijeikis et al. [160] designed a U-Net-like network that uses MobileNet V2 as an encoder for spatial feature extraction. Then, LSTM was used for temporal feature extraction, followed by two FCLs for violence detection. In 2023, Qasim and Verdu [161] proposed an abnormal behavior detection method that integrates ResNet-50 to learn high-level feature representations and SRU to capture temporal features. In the same year, Chaurasia and Jaiswal [117] combined U-Net network and Conv-LSTM to handle spatial information and temporal motion. First, bilateral filtering was applied to obtain smoothed frames. Then, a combination of U-Net architecture and Conv-LSTM was utilized to capture spatial information and temporal motion, and generate future frames. Finally, an anomaly score was computed based on the peak signal-to-noise ratio and the structural similarity index measure. In 2024, Vosta and Yow [86] combined ResNet 50 and Conv-LSTM with a multi-head self-attention layer for violent behavior detection.
- **AEs + LSTM-based methods** In 2018 Yang et al.[162] developed a two-stream framework for abnormal pedestrian behavior detection and localization. Two networks that combined Convolutional Autoencoder (Conv-AE) with Conv-LSTM are used to capture spatial and temporal variance and reconstruct raw data and optical flow of normal behaviors. The designed networks adopted a weighted eclidian loss to minimize the impact of the complex background and focus only on the moving foreground. Then a global-local analysis was utilized for anomaly detection and localization. In 2019, Nawaratne et al. [163] presented an unsupervised approach based on continuous learning. The authors used a spatio-temporal AE to learn normal behaviors, integrating

CNN to get spatial information and Conv-LSTM to capture temporal dynamics. The classified results were validated by a human observer and then used to update normal behavior through fuzzy aggregation. This updated normal behavior was fed back into the model to update the previously learned knowledge. In 2021, Li et al. [164] proposed a two stream based framework composed of a spatial stream and a temporal stream. The spatial stream fed RGB frames into a deep spatiotemporal AE to extract appearance characteristics, while the temporal stream fed optical flow frames into the same network to extract motion patterns. The two stream were fused using joint reconstruction error to extract spatiotemporal characteristics for anomaly detection. In 2022, Pawar and Attar [165] proposed an approach for anomaly detection and localization based on pipelined AEs and One-Class Learning. The authors used Conv-AE and Seq2sp LSTM AE [166] to learn appearance and temporal features and applied multivariate Gaussian distribution for anomaly identification. In 2023, Cheng et al. [167] introduced a framework that integrates motion and appearance branches for feature frame prediction. For the appearance branch, an encoder was used to capture appearance information from consecutive frames. The motion branch employed dynamic graph recurrent neural network based on LSTM layers to extract motion information from optical flow frames. An attention module was then applied to fuse spatial and temporal information. Finally, the decoder predicts the future frame and compute an anomaly score based on the predicted frames to detect anomalies. In 2024, Aslam and Kolekar [168] proposed an anomaly detection model that combines the benefits of ConvAE, LSTM encoder–decoder and attention mechanisms. The model encoded spatial information from normal videos and then captured temporal features using a ConvLSTM layer. An attention-based decoder was used to focus on relevant temporal features by computing alignment scores and attention weights from the encoder and decoder hidden states.

- **ST-GCN-based methods** In 2021, Luo et al. [169] employed AlphaPose as a human pose estimator to extract skeletons from each video frame. Next, stacked ST-GCNs [170] were used to model the spatio-temporal relationships of the extracted skeletons. The output of the ST-GCNs was then passed to a prediction module that integrates an FCL to predict the future skeleton. In 2022, Yang et al. [171] proposed a two-stream framework for abnormal event detection. First, YOLOv3 was used to detect objects from video frames. In the first stream, frames with detected pedestrians were processed by Alpha-Pose to estimate body joints. ST-GCN was then applied to capture the spatio-temporal features of the normal body joints, followed by an FCL to output their normality scores. The second stream computed the normality scores of other objects. Finally, the outputs from both streams were fused at the decision level to generate the final normality score. In 2023, Fei et al. [6] developed a two stream network which incorporates the optical flow and human pose information for fall detection. The first stream used optical flow and VGG-16 to learn motion features. For the second stream, the joints of the human body were detected and passed through a GCN-based network to learn appearance features. The outputs of the two streams were concatenated and then fed into a classifier to generate the classification result.
- **GAN + Optical flow-based methods** In 2017, Ravanbakhsh et al. [172] introduced two conditional GANs for image translation, which are trained on normal frames along with their corresponding optical flows to model normal patterns. In 2018, Liu et al. [173] proposed a video frame prediction network based on the U-Net architecture. They applied intensity, gradient, and optical flow aiming to enhance the generated image and enforce its closeness to the ground truth. In 2019, Ravanbakhsh et al. [174] proposed an adversarial training based-approach for abnormal crowd behavior detection. The proposed approach adopted a cross-channel notwork with two-channel GAN to generate video frames and its optical flow graphs. In 2022, Alafif et al. [48] proposed an AHB detection framework that combines optical flow and GAN. A fine-tuned GAN was used to reconstruct optical flow maps corresponding to normal frames and distinguish between real and generated optical flow graphs. In 2024, Ehsan et al. [175] presented an unsupervised framework for violent behavior detection. A YOLO network was used for person detection, followed by the Farneback method [176] for motion features extraction. The obtained result was fed into a spatiol-temporal network that integrates a U-Net and PatchGan [177] for frame reconstruction and human behavior classification.

- **Semantic sequential frame-based methods**
  - **Transformers-based methods** In 2022, Li et al. [178] implemented a multi-sequence learning approach based on the transformer architecture.

The designed approach was trained to learn both video-level anomaly probability and snippet-level anomaly scores for weakly supervised video anomaly detection. In 2023, Rendón-Segador et al. [179], proposed a method that combined adversarial neural structured learning(NSL) with VideoSwin [180] to recognize violent activities in videos. In the same year, Joo et al. [181] proposed to use the Vision Transformer (ViT) encoder of the CLIP [182] model to extract discriminative representations. Each video was divided into a set of frame snippets, and the ViT encoder was then applied to extract ViT features from these snippets. Temporal self-attention was subsequently used to obtain anomaly attention features. These features were then passed through a difference maximization trainer to weakly train the anomaly classifier. In 2024, Zhu et al. [183], introduced an approach that combines GCN and Transformer to model normal patterns of human behaviors. First, human 2D skeletons was extracted and tracked using Alpha-Pose and YOLOx-x, then processed into sequences using a sliding window. A Space-Time Graph-Transformer Encoder generated latent vectors to learn normal behavior features stored in memory during training. Finally, skeleton sequences were reconstructed and predicted based on this memory-enhanced information, and anomaly scores were calculated by fusing reconstruction and prediction errors.

- **VLM-based methods** In 2024, Wu et al. [184] developed a dual-branch framework that leverages both the visual and textual encoders of the frozen CLIP model. The first branch employs visual features for coarse-grained binary classification, while the second aligns visual and textual modalities to enable fine-grained anomaly detection. To model temporal relationships, the authors introduced a Local-Global Temporal Adapter (LGT-Adapter), composed of two components. A Local Temporal Adapter, based on a transformer encoder, captures local temporal dependencies, and a Global Temporal Adapter, a lightweight GCN module designed to model global temporal dependencies.
- **VLM + LLM-based methods** In 2024, Zanella et al. [185] were the first to propose a training-free video anomaly detection method that leverage LLMs for anomaly scoring with modality-aligned VLMs. First, they employed BLIP2 [186], a VLM-based captioning model, to generate textual descriptions for each video frame. Then, they introduced an image-text caption cleaning component to refine noisy descriptions based on cross-modal similarity. Next, the cleaned frame captions were aggregated within temporal windows, and LLaMA [187] was prompted to generate temporal summaries and estimate anomaly scores based on these summaries. Finally, the anomaly scores were refined using a Video-Text Score Refinement component based on video-text similarity.

4.4.5 Model output

The approaches proposed in the reviewed studies mainly focus on three key tasks: detection, recognition, and localization of anomalies. Detection aims to determine the presence or absence of the anomaly in the video. Recognition involves classifying the detected anomaly into a predefined class, which is crucial for analyzing and understanding the nature of the anomaly. Localization focuses on identifying the exact area where the anomaly occurs within the video frame, allowing for a more detailed and targeted analysis of the the detected anomaly. Our findings show that detection is the most frequently addressed task, followed by localization and recognition. It is important to note that a simple detection of the presence of the anomaly is often insufficient, especially in contexts that involve human life. The recognition of the type of anomaly and the identification of its exact location are essential in many contexts as they help make rapid and suitable decisions.

In addition, we identified different levels of granularity in video analysis, including pixel-level, frame-level, video-level, and patch-level. The pixel level focuses on identifying anomalies at the individual pixel scale, providing a fine-grained analysis of each pixel within the frame. The frame level involves the identification of the anomaly within each frame of the video. The video level provides a global analysis of the entire video, representing a low-level of granularity. The patch level refers to splitting the frame or video into patches, and detecting anomalies in these patches rather than in the entire frame or video.

4.4.6 Model evaluation

There are two methods generally used to evaluate the performance of deep learning models: quantitative evaluation and human-based evaluation. Quantitative evaluation consists of using predefined evaluation metrics to measure the model's performance. Human-based evaluation involves incorporating human feedback into the evaluation process, where an expert in the domain analyzes the results generated by the model and makes the final decision. Our findings show that the majority of the reviewed studies rely only on the quantitative evaluation. While the use of performance metrics is the main part of the evaluation process,

incorporating human feedback remains critical to validate the applicability of automatically generated results in real world scenarios, especially when human lives are involved.

4.5 What computing paradigms and techniques are used to enable real-time AHB detection? (RQ5)

The detection of AHB in video surveillance is a time-sensitive application that requires rapid processing and immediate response to act on any potential danger. To achieve real-time detection, various computing paradigms have been adopted, including cloud computing, edge computing, and terminal computing.

In cloud computing, data are initially captured on the user device and then transmitted to cloud centers for processing. A notable example is Alia et al. [41]. The authors proposed a real-time pushing detection framework that adapts EfficientNetV2B0 [188] and integrates it with GPU-based RAFT [74], the wheel color method and live camera technology on a cloud platform. However, this computing paradigm typically requires high bandwidth and has a long response time due to network latency, making it unsuitable for time-sensitive applications [189].

Edge computing refers to processing data on edge nodes located near the data sources thereby reducing network latency and enabling rapid response times [190]. Similarly, terminal computing, also called user devices' computing, involves processing data directly on the user device itself (cameras, drones, smartphones, computers), ensuring very low response times [191]. These two paradigms are characterized by limited computational power, memory and energy availability. Deploying deep learning models on resource-constrained devices presents a significant challenge, as these models are extremely resource-consuming during both training and inference phases.

Various approaches used lightweight models that are less computationally intensive and can easily be deployed. Ullah et al. [153] employed the Mobile-SSD CNN model to provide real-time person detection in Closed-Circuit TeleVision (CCTV) camera streams. Cheng et al. [192] proposed a secure and lightweight video anomaly detection framework for edge computing environments that integrates Conv-AE and Conv-LSTM networks with privacy-preserving techniques. A fine-grained Bloom filter-based access control policy was introduced to authenticate legitimate users, without lacking the privacy of raw personal attributes. Chang et al. [193] developed a fall detection method that operates on edge computing environment. The proposed method employed OpenPose-light [194] to detect human joints and used LSTM to perform fall recognition. Doshi and Yilmaz [105] employed YOLOv3 as a lightweight object detection model to extract relevant features. Mehmood, [195] adopted a low computational cost method, stacked grayscale 3-channel images (SG3Is) [196], for modeling motion features. They also introduced a lightweight CNN architecture for abnormal behavior detection. In [31], Hussain et al, developed a lightweight 3DCNN network for abnormal activity recognition in CCTV surveillance systems. Gallo et al. [125] proposed a smart system that leverages edge computing and a fine-tuned YOLOv4-tiny [197] to detect personal protective equipment in real time within industrial environments.

Lightweight systems often compromise performance to obtain smaller and faster architectures usually leading to accuracy degradation especially in complex scenarios. Quantization is also considered as an effective way to integrate deep learning models into resource-constrained devices. In deep learning quantization, the weights and activation tensors are stored in lower bit precision than the precision they are usually trained in [198]. However, lower precision quantization may introduce noise to the model that could lead to a drop in accuracy. An hybrid edge–cloud architecture can be adopted to address this limitation. A notable example is Ullah et al. [199]. The authors proposed a two-stream architecture which the first stream enables instant anomaly detection at the edge using a lightweight CNN, while the second stream performs detailed anomaly identification in the cloud using BD-LSTM. A further example is presented by Rajavel et al. [56]. The authors proposed an IoT-based smart healthcare video surveillance system that leverages edge and cloud computing paradigms to reduce network bandwidth and response time. The system integrates background subtraction and CNN to detect and classify abnormal falling activities of patients and elderly individuals. Furthermore, federated learning [200] presents a promising solution by enabling collaborative model training across distributed, resource-constrained devices thereby preserving robust performance while ensuring low-latency responses.

4.6 What are the commonly used datasets to train and evaluate models? (RQ6)

Through our systematic analysis, we identified 66 datasets employed in the experiments of the reviewed studies. We found that 67% of these datasets are publicly accessible, while 33% remain unavailable. Table 5, provides a comparison of the most popular datasets, based on key characteristics including category, total number of frames, resolution, scene type, scene density, variations, and examples of anomalies. Few available datasets offer high-resolution video frames, which are essential for detecting subtle abnormal behaviors. Some examples of such datasets are IITB-Corridor [201] (1920×1080), Street Scene [202] (1280×720), and AIRTLab [122] (1920×1080).

Table 5 Comparison of datasets for abnormal behavior

Dataset	Category	#Frames	Resolution	Scene		Variations	Anomalies
				Type	Density		
UCF-crime [40]	AB	13M	Variable	Hybrid	Crowd	Yes	Assault, fighting, robbery, etc
UMN [214]	SB	7710	320 × 240	Hybrid	Crowd	No	Escaping
Violent Flow [216]	AB	22,156	320×240	Outdoor	Crowd	Yes	Violence
Subway Entrance [215]	SB	136,524	512×384	Indoor	Crowd	No	Wrong direction,loitering, etc
Subway Exist [215]	SB	72,401	512×384	Indoor	Crowd	No	Wrong direction, no payment, etc
Le2i Fall Detection [206]	H-RB	7832	320×240	Indoor	Uncrowded	No	Fall events
Hockey Fight [212]	AB	50000	720×576	Indoor	Uncrowded	No	Fights
RWF-2000 [213]	AB	300,000	Variable	Hybrid	Crowd	Yes	Fights
UCSD Ped1 [207]	SB	14,000	238×158	Outdoor	Uncrowded	No	Walking on grass, non pedestrian entities
UCSD Ped2 [207]	SB	4,560	360×240	Outdoor	Uncrowded	No	Walking on grass, non pedestrian entities
CUHK Avenue [208]	AB, SB	30,652	640×360	Outdoor	Uncrowded	No	Wrong direction, Run, throw, etc
URFD [205]	H-RB	-	640 × 480	Indoor	Uncrowded	No	Fall events
Driver_data [54]	DB	-	640×360	Indoor	Uncrowded	No	Eating, drinking, using the phone
ShanghaiTech [209]	AB, SB	317,398	856×480	Outdoor	Hybrid	No	Chasing, jumping, fighting, etc
IITB-Corridor [201]	AB, SP	483,566	1920×1080	Outdoor	Crowd	No	Fights, hiding face, suspicious objects
Movies [212]	AB	6000	360×250	Hybrid	Uncrowded	Yes	Fights
FDD [203]	H-RB	108,476	-	Indoor	Uncrowded	No	Fall events
Multicam [204]	H-RB	261,137	720x480	Indoor	Uncrowded	Yes	Fall events
Street Scene [202]	AB, DB	203,257	1280x720	Outdoor	Uncrowded	Yes	Jaywalking, illegally parked, car u-turn
XD-Violence [210]	AB	-	Variable	Variable	Variable	Yes	Shooting, fights, riot, etc
PETS2009 [217]	SB	42182	Variable	Outdoor	Crowd	Yes	Run, panic
Behave [219]	AB, SB, DB	200,000	640×480	Outdoor	Uncrowded	Yes	Chasing, fighting, following a person, etc
AIRTLab [122]	AB	59,115	1920x1080	indoor	Uncrowded	No	Kicks, punches, slaps, etc
MDV [211]	AB, SB, DB	-	960×540	Outdoor	Variable	Yes	Stealing, loitering, improper parking, etc
IITH accident [218]	DB	990,138	-	Outdoor	Crowd	Yes	Accidents
UT-interaction [220]	AB	36,000	720x480	Outdoor	Uncrowded	YES	Push, kick, punch and point

We compare publicly available datasets commonly used for AHB detection highlighting key characteristics, including category, total number of frames, resolution, scene type, scene density, variations, and examples of anomalies. Category: Aggressive Behavior(AB), Suspicious Behavior(SB), Driver Behavior(DB), and Health-Related Behavior(H-RB). Variations: Yes, if the dataset include environmental variations, otherwise, No

In addition, our findings suggest that the majority of the datasets fall within the categories of aggressive behavior and suspicious behavior. There is a notable lack of publicly available datasets related to driver behavior and health-related behavior, which is a significant gap in the current literature. Moreover, we observe that datasets classified as health-related behavior, such as FDD [203], Multicam[204], URFD [205], and Le2i Fall Detection [206], primarily focus on fall-related events, lacking representation of other health-related anomalies such as repetitive movements, abnormal gait, and self-harm behavior. We highlight the importance of having more diverse data that cover a wider range of health-related behaviors.

Notably, several datasets including UCF-Crime [40], UCSD [207], CUHK Avenue [208], ShanghaiTech [209], XD-Violence [210], MDV [211], and Driver_data [54] cover various types of abnormal behavior. In contrast, other datasets such as Hockey Fight [212], RWF-2000 [213], UMN [214], Subway Entrance and Exist [215], Violent Flow [216], PETS2009 [217], Movies [212], and IITH accident [218] are designed to detect specific types of behavior. We emphasize the need for more datasets that include different types of anomalies. Such datasets are essential for training models that can generalize effectively in real-world scenarios.

Our analysis also reveals that popular benchmark datasets such as AIRTLab [122], UMN [214], UCSD [207] and CUHK Avenue [208] do not consider environmental variations such as lighting and weather conditions. Furthermore, datasets such as UCF-Crime [40] (13M frames), IITB-Corridor [201] (483,566 frames), and Behave [219] (200,000 frames), provide a large amount of data, which is essential for training deep learning models. In contrast, smaller datasets such as UCSD Ped2 [207] (4,560 frames), Movies [212] (6,000 frames), and UMN [214] (7,710 frames) often lead to poor generalization when used for model training.

4.7 What are the performance metrics considered for model evaluation? (RQ7)

Performance metrics are essential tools for measuring the performance of deep learning models and assessing its effectiveness. These metrics provide quantitative measures that can identify the strengths and weaknesses of different approaches which help to compare them and determine the most suitable one for a given task. We found that Area Under the Curve (AUC), Accuracy(ACC), Equal Error Rate(EER), Precision, Recall, F1-score, and False Alarm Rate (FAR) are most frequently applied metrics. We observed that all identified papers used either AUC, ACC or both together as their main evaluation metrics. We also noticed that Precision and Recall are often used together. It is important to note that these metrics can also be applied in various other tasks, such as text classification and time-series data analysis. There is a significant lack of metrics specifically designed for visual-based systems. In Table 6, we summarize the performance evaluation of the studies discussed in Sections 4.4.4 and 4.5

Table 6 Summary of performance evaluation results

Paper	Used datasets				Values (%)		
	UCF-Crime	UCSD Ped2	URFD	Driven-Data	AUC	ACC	EER
Sabokrou et al. [92]		✓					11
Lalit et al. [76]		✓			97.9		9
Zhou et al. [156]		✓			86		24.4
Sultani et al. [40]	✓				75.41		
Pang et al. [149]		✓			83.2		
Zhang et al. [54]				✓		81.91	
Ullan et al. [159]	✓				85.53	85.53	
Huszar et al. [43]	✓				97.1	90.1	
Qasim and Verdu [161]	✓				91.64	91.44	
Vosta and Yow [86]	✓				97.48	92.98	
Yang et al. [171]		✓			92.7		
Fei et al. [6]			✓			100	
Mehmood [195]			✓		98.71	98.86	
Chang et al. [193]			✓			97.6	
Doshi and Yilmaz [157]		✓			97.8		
Doshi and Yilmaz [105]		✓			97.2		
Sabokrou et al. [221]		✓			99.6		8.2
Nguyen and Meunier [222]		✓			96.2		
Wu et al. [99]		✓			92.8		12.5
Li et al. [154]		✓			95.1		
Hao et al. [155]		✓			96.9		8
Kamoona et al. [223]	✓				79.49		
Yang et al. [162]		✓			94.8		12
Nawaratne et al. [163]		✓			91.1		8.9
Li et al. [164]		✓			97.2		6.9
Cheng et al. [167]		✓			97.0		
Aslam and Kolekar [168]		✓			96.2		11.2
Ravanbakhsh et al. [172]		✓			93.5		14
Liu et al. [173]		✓			95.4		
Ravanbakhsh et al. [174]		✓			95.5		11
Shin et al. [137]		✓			93		11
Chen et al. [150]		✓			96.3		6
Alafif et al. [48]		✓			95.7		
Chaurasia and Jaiswal [117]		✓			97.1		8.1
Ehsan et al. [175]		✓			93		
Li et al. [178]	✓				85.62		
Rendón-Segador et al. [179]	✓				99.98		
Joo et al. [181]	✓				87.58		
Wu et al. [184]	✓				82.45		
Zanella et al. [185]	✓				80.28		

for which results are available. Specifically, we highlight their performance using AUC, ACC, and EER, as these are the most commonly used metrics across the four benchmark datasets: UCF-Crime, UCSD Ped2, URFD, and Driver-data.

4.8 What future research directions can address the current limitations of AHB recognition systems and enhance their performance in real-world environments? (RQ8)

Despite the significant advancements made in this field, more research are still required to address the challenges encountered in real-world scenarios. In this section, we highlight some research directions that could be explored in the future.

Long-term temporal modeling The detection of anomalies in videos largely depends on a model's ability to capture both short-term and long-term temporal dependencies within video sequences. Despite the significant progress achieved with deep learning, many models still struggle to effectively capture long-term temporal dynamics within video data. Future research could focus on improving temporal modeling capabilities by integrating advanced models such as transformers [184, 224].

Multi-modal learning The rise of multimodal models has transformed the landscape of deep learning, enabling richer and more context-aware representations across modalities. Models such as Contrastive Language–Image Pre-training (CLIP) [182] have demonstrated a remarkable ability to align visual and textual representations within a shared semantic space, enabling semantic detection of abnormal behaviors based on simple textual descriptions. Similarly, Imagebind [225] learns a joint embedding across six different modalities- images, text, audio, depth, thermal, and IMU data– using images as the binding modality. More effort is needed to explore models specifically tailored for handling multiple modalities, enabling more intelligent and context-aware systems.

Open-World approaches The real world is characterized by continuous evolution and variability. Consequently, it is unrealistic to assume that all abnormal classes expected during inference are already seen during training [226]. Therefore, rather than relying on a closed-world model limited to a predefined set of anomaly classes [149, 156], it is more effective to adopt an open-world model capable of recognizing previously unseen anomalies and adapting to new scenarios. Despite its significance for real-world applications, this open-world perspective remains relatively under-explored in the existing literature, highlighting an important direction for future research.

Training-free strategy Training-free approaches use pre-trained models to detect anomalies without requiring additional training process. Serval advanced models such as CLIP [182], BLIP-2 [186], LLaMa [187], Mistral-7B [227] and GPT-4V [228] have demonstrated exceptional capabilities in learning generalized representations and performing various tasks through prompt engineering or zero-shot learning without task-specific fine-tuning [185, 229]. More research on training-free strategies is expected, especially in data-scarce domains.

Explainability and transparency Explainability has emerged as a critical aspect in video anomaly detection, especially in sensitive and high-stakes environments [230]. An explainable model focuses not only on identifying anomalies but also on providing transparent and understandable explanations for its decisions. We expect more explainable and transparent models in this field in the future.

5 Comparison with existing SLRs

In this section, we present a comparative analysis of our work with the four previously identified SLRs [24–27]. Table 7 illustrates this comparison based on some key properties such as time range covered, reviewed papers, databases consulted, research questions posed, main focus and limitations.

Omarov et al. [24] discussed the use of machine learning, SVM and deep learning models for violence detection. A total of 80 studies published between 2015 and 2021 were

Table 7 Comparison with existing SLRs

Ref	Time range	Studies	DB	RQs	Main focus	Limitations
Omarov et al. [24]	2015-2021	80	5	4	Machine learning, SVM, and deep learning for violence detection	Focus on particular deep learning models
Bouhsissin et al. [25]	2015-2022	93	5	6	Abnormal driver behavior classification using machine learning, deep learning and statistical techniques	Lack information on data fusion-based systems
Samaila et al. [26]	2003-2023	530	1	7	Anomaly detection using classical machine learning, and deep learning methods	No quantitative meta-analysis provided
Gaya-Morey et al. [27]	2019-2023	151	5	5	Deep learning for fall detection and human activity recognition for the elderly	Ignore earlier relevant deep learning-based work.
Ours	2016-2024	140	7	8	Deep learning for AHB recognition	Lack an in-depth discussion of multi-modal techniques and datasets

reviewed, which is a limited number compared to the volume of publications available during that period. In addition, the authors did not cover the full range of deep learning models employed during the selected period such as AEs, GCN and GAN. Their focus was limited to CNN and LSTM-based approaches. Bouhsissin et al. [25] explored driver behavior classification using machine learning, deep learning and statistical techniques. The authors reviewed 93 research papers published between 2015 and 2022. Although the study discussed the different data sources used to extract data and classify driver behavior such as camera, GPS and smartphone sensors, accelerometer, and gyroscope, it lacks information on data fusion-based systems. Samaila et al. [26] addressed abnormal human action recognition using traditional, classical and deep learning techniques. The authors examined studies published between 2003 and 2023 providing an analysis of 530 papers. Although the review covers a large number of publications, it lacked a quantitative meta-analysis of the performance of the reviewed approaches. Gaya-Morey et al. [27] summarized the advancement in deep learning-based fall detection and human activity recognition for the elderly. However, the study included only research paper published between 2019 and 2023, thereby excluding earlier relevant deep learning-based work. In addition, the review focuses on the use of specific visual data, including RGB, depth, and infrared, while overlooking other data modalities that can be crucial for detecting abnormal activities among the elderly.

In contrast, our work provides a broader perspective on the use of deep learning for video-based AHB recognition. We offer a detailed analysis of 140 studies published between 2015 and 2023. While previous research typically focuses on limited types of anomalies, our study addresses a wide range of abnormal behavior. Furthermore, we include an extensive review of deep learning models employed in the field, considered advanced and pre-trained architectures. In addition, the study focuses on essential aspects such as real-time detection, datasets, and future research directions. However, our work primarily focuses on visual-based methods and lacks an in-depth discussion of multi-modal techniques and datasets that incorporate various data sources such as video frames, audio, and text.

6 Conclusion

Although intelligent surveillance cameras are widely implemented in various areas of daily life, developing an efficient and robust system that automatically analyzes human behaviors and detect anomalies using traditional machine learning has many limitations. Deep learning has emerged as an effective solution, providing advanced architectures for extracting meaningful information and capture complex patterns from videos. In this work, a systematic review of recent advancements in deep learning for video-based AHB recognition has been presented. Following a predefined methodology, we have executed four search queries on online academic databases to collect relevant research studies. After the screening process, we have identified 140 studies, including 118 journal articles and 22 conference papers, published between 2016 and 2024. These works were thoroughly examined to address the eight RQs raised in this SLR. Our analysis has shown that there are many types of AHB. We have organized them into four categories: aggressive behavior, suspicious behavior, drive behavior and health-related behavior. Then, we have explored the deep learning models used in the reviewed studies. We have found that 86% of the identified studies employed only deep learning models, while 14% combined deep learning with traditional machine learning. CNN, AEs, and RNN, along with their variants, are among the most widely used models. We have also explored pre-trained models and found that VGGNet, ResNet, and Inception are the most frequently employed. After that, we have discussed the key steps of a video-based AHB recognition system and have identified the commonly used methods at each step. The approaches proposed in the reviewed studies were examined and classified into three main categories: Frame-based methods, sequential frame-based methods, and semantic sequential frame-based methods. The computing paradigms and techniques used to enable real-time detection were also discussed. In fact, deploying deep learning models on resource-constrained devices such as edge or local devices is still challenging, as these models are extremely resource-consuming during both training and inference phases. Lightweight models are very adopted as an alternative. Furthermore, this review has analyzed the most commonly used datasets, discussing their limitations. We have observed that only a few datasets have high-resolution video frames such as IITB-Corridor, Street Scene, and AIRLab. We have also noticed that popular datasets including AIRLAB, UMN, UCSD, and CUHK Avenue, omit environmental and climatic conditions. Another important limitation is the lack of datasets related to driver and health behaviors. The study has also examined the performance metrics used for model evaluation and has highlighted the need for specific metrics related to the domain. For future research directions, we expect more research on temporal modeling, multi-modal fusion, open-world recognition, training-free strategy, and explainability and transparency. Finally, we have provided a comparative analysis of our SLR with existing works based on five key criteria: time range, number of studies reviewed, number of databases consulted, number of research questions posed, and main focus. Given the rapid evolution of

deep learning models, there is a growing need for such a study that reviews the latest advancements of these models in video-based AHB recognition.

Data Availability No data was generated for the research described in the article.

Declarations

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Hu K, Yang H, Jin Y, Liu J, Chen Y, Zhang M, Wang F (2023) Understanding user behavior in volumetric video watching: Dataset analysis and prediction. In: Proceedings of the 31st ACM international conference on multimedia pp 1108–1116
2. Pareek P, Thakkar A (2021) A survey on video-based human action recognition: Recent updates datasets challenges and applications. Artif Intell Rev 54(3):2259–2322
3. Dhiman C, Vishwakarma DK (2019) A review of state-of-the-art techniques for abnormal human activity recognition. Eng Appl Artif Intell 77:21–45
4. Chaturvedi K, Dhiman C, Vishwakarma DK (2024) Fight detection with spatial and channel wise attention-based convlstm model. Expert. Syst. 41(1):13474
5. Ansari MA, Singh DK (2022) An expert video surveillance system to identify and mitigate shoplifting in megastores. Multimed Tools Appl 1–29
6. Fei K, Wang C, Zhang J, Liu Y, Xie X, Tu Z (2023) Flow-pose net: An effective two-stream network for fall detection. Vis Comput 39(6):2305–2320
7. Khairdoost N (2022) Driver behavior analysis based on real on-road driving data in the design of advanced driving assistance systems. PhD thesis The University of Western Ontario (Canada)
8. Iqbal N, Saad Missen MM, Salamat N, Prasath VS (2019) On video based human abnormal activity detection with histogram of oriented gradients. Handb Multimed In Secur Tech App 431–448
9. Iqbal JM, Lavanya J, Arun S (2015) Abnormal human activity recognition using scale invariant feature transform. Int J Curr Eng Technol 5(6):3748–3751
10. Ke Y, Sukthankar R, Hebert M (2010) Volumetric features for video event detection. Int J Comput Vision 88:339–362
11. Colque RVHM, Caetano C, Andrade MTL, Schwartz WR (2016) Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. IEEE Trans. Circuits Syst. Video Technol 27(3):673–682
12. Al-Nawashi M, Al-Hazaimeh OM, Saraee M (2017) A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. Neural Comput Appl 28:565–572
13. Yao C, Su X, Wang X, Kang X, Zhang J (2021) Ren J (2021) Motion direction inconsistency-based fight detection for multiview surveillance videos. Wirel Commun Mob Comput 1:9965781
14. Saligrama V, Chen Z (2012) Video anomaly detection based on local statistical aggregates. In: 2012 IEEE conference on computer vision and pattern recognition, IEEE, pp 2112–2119
15. Mengist W, Soromessa T, Legese G (2020) Method for conducting systematic literature review and meta-analysis for environmental science research. MethodsX 7:100777
16. Nayak R, Pati UC, Das SK (2021) A comprehensive review on deep learning-based methods for video anomaly detection. Image Vis Comput 106:104078
17. Patrikar DR, Parate MR (2022) Anomaly detection using edge computing in video surveillance system. Int J Multimed Inf Retr 11(2):85–110
18. Berroukham A, Housni K, Lahraichi M, Boulfrifi I (2023) Deep learning-based methods for anomaly detection in video surveillance: A review. Bull Electr Eng Inf 12(1):314–327
19. Duong HT, Le VT, Hoang VT (2023) Deep learning-based anomaly detection in video surveillance: A survey. Sensors 23(11):5024
20. Huang J, Yakun C, Tingting S (2023) Investigating of deep learning-based approaches for anomaly detection in iot surveillance systems. Int J Adv Comput Sci Appl. 14(12)
21. Liu Y, Yang D, Wang Y, Liu J, Liu J, Boukerche A, Sun P, Song L (2024) Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. ACM Comput Surv 56(7):1–38
22. Wastupranata LM, Kong SG, Wang L (2024) Deep learning for abnormal human behavior detection in surveillance videos–a survey. Electron 13(13):2579
23. Negre P, Alonso RS, González-Briones A, Prieto J, Rodríguez-González S (2024) Literature review of deep-learning-based detection of violence in video. Sensors 24(12):4016
24. Omarov B, Narynov S, Zhumanov Z, Gumar A, Khassanova M (2022) State-of-the-art violence detection techniques in video surveillance security systems: A systematic review. PeerJ Comput Sci 8:920
25. Bouhsissin S, Sael N, Benabbou F (2023) Driver behavior classification: A systematic literature review. IEEE Access
26. Samaila YA, Sebastian P, Singh NSS, Shuaibu AN, Ali SSA, Amosa TI, Abro GEM, Shuaibu I (2024) Video anomaly detection: A systematic review of issues and prospects. Neurocomputing 127726
27. Gaya-Morey FX, Manresa-Yee C, Buades-Rubio JM (2024) Deep learning for computer vision based activity recognition and fall detection of the elderly: A systematic review. Appl Intell 1–26
28. Kitchenham B (2004) Procedures for performing systematic reviews. Keele UK Keele Univ 33(2004):1–26
29. Mohtavipour SM, Saeidi M, Arabsorkhi A (2022) A multi-stream cnn for deep violence detection in video sequences using handcrafted features. Vis Comput 38(6):2057–2072
30. Ullah FUM, Obaidat MS, Muhammad K, Ullah A, Baik SW, Cuzzolin F, Rodrigues JJ, Albuquerque VHC (2022) An intelligent system for complex violence pattern analysis and detection. Int J Intell Syst 37(12):10400–10422
31. Hussain A, Ullah H, Ullah A, Imran AS, Lee My, Rho S, Sajjad M et al (2021) Anomaly based camera prioritization in large scale surveillance networks
32. Asad M, Yang Z, Khan Z, Yang J, He X (2019) Feature fusion based deep spatiotemporal model for violence detection in videos. In: Neural information processing: 26th international conference ICONIP 2019 Sydney NSW Australia December 12–15 2019 Proceedings Part I 26, Springer, pp 405–417
33. Serrano I, Deniz O, Espinosa-Aranda JL, Bueno G (2018) Fight recognition in video using hough forests and 2d convolutional neural network. IEEE Trans Image Process 27(10):4787–4797
34. Asad M, Yang J, He J, Shamsolmoali P, He X (2021) Multi-frame feature-fusion-based model for violence detection. Vis Comput 37(6):1415–1431
35. Fenil E, Manogaran G, Vivekananda G, Thanjaivadivel T, Jeeva S, Ahilan A et al (2019) Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional lstm. Comput Netw 151:191–200

36. Irfanullah Hussain T, Iqbal A, Yang B, Hussain A (2022) Real time violence detection in surveillance videos using convolutional neural networks. Multimed Tools Appl 81(26):38151–38173
37. Magdy M, Fakhr MW, Maghraby FA (2023) Violence 4d: Violence detection in surveillance using 4d convolutional neural networks. IET Comput Vision 17(3):282–294
38. Garcia-Cobo G, SanMiguel JC (2023) Human skeletons and change detection for efficient violence detection in surveillance videos. Comput Vis Image Underst 233:103739
39. Vrskova R, Hudec R, Kamencay P, Sykora P (2022) A new approach for abnormal human activities recognition based on convlstm architecture. Sensors 22(8):2946
40. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6479–6488
41. Alia A, Maree M, Chraibi M, Toma A, Seyfried A (2023) A cloud-based deep learning framework for early detection of pushing at crowded event entrances. IEEE Access
42. Buttar AM, Bano M, Akbar MA, Alabrah A, Gumaei AH (2023) Toward trustworthy human suspicious activity detection from surveillance videos using deep learning. Soft Comput 1–13
43. Huszar VD, Adhikarla VK, Négyesi I, Krasznay C (2023) Toward fast and accurate violence detection for automated video surveillance applications. IEEE Access 11:18772–18793
44. Khan SU, Haq IU, Rho S, Baik SW, Lee MY (2019) Cover the violence: A novel deep-learning-based approach towards violence-detection in movies. Appl Sci 9(22):4963
45. Sudhakaran S, Lanz O (2017) Learning to detect violent videos using convolutional long short-term memory. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS), IEEE, pp 1–6
46. Bala A, Kaushal R (2023) Jaywalking detection and localization in street scene videos using fine-tuned convolutional neural networks. Multimed Tools Appl 82(22):34771–34791
47. Martínez-Mascorro GA, Abreu-Pederzini JR, Ortiz-Bayliss JC, Garcia-Collantes A, Terashima-Marín H (2021) Criminal intention detection at early stages of shoplifting cases by using 3d convolutional neural networks. Computat 9(2):24
48. Alafif T, Alzahrani B, Cao Y, Alotaibi R, Barnawi A, Chen M (2022) Generative adversarial network based abnormal behavior detection in massive crowd videos: A hajj case study. J Ambient Intell Humaniz Comput 13(8):4077–4088
49. Direkoglu C (2020) Abnormal crowd behavior detection using motion information images and convolutional neural networks. IEEE Access 8:80408–80416
50. Mehmood A (2021) Abnormal behavior detection in uncrowded videos with two-stream 3d convolutional neural networks. Appl Sci 11(8):3523
51. Celaya-Padilla JM, Galván-Tejada CE, Lozano-Aguilar JSA, Zanella-Calzada LA, Luna-García H, Galván-Tejada JI, Gamboa-Rosales NK, Velez Rodriguez A, Gamboa-Rosales H (2019) “Texting & driving" detection using deep convolutional neural networks. Appl Sci 9(15):2962
52. Huang W, Liu X, Luo M, Zhang P, Wang W, Wang J (2019) Video-based abnormal driving behavior detection via deep learning fusions. IEEE Access 7:64571–64582
53. Huang S, Yang L, Chen W, Tao T, Zhang B (2021) A specific perspective: Subway driver behaviour recognition using cnn and time-series diagram. IET Intel Transport Syst 15(3):387–395
54. Zhang C, Li R, Kim W, Yoon D, Patras P (2020) Driver behavior recognition via interwoven deep convolutional neural nets with multi-stream inputs. IEEE Access 8:191138–191151
55. Chhetri S, Alsadoon A, Al-Dala’in T, Prasad P, Rashid TA, Maag A (2021) Deep learning for vision-based fall detection system: Enhanced optical dynamic flow. Comput Intell 37(1):578–595
56. Rajavel R, Ravichandran SK, Harimoorthy K, Nagappan P, Gobichettipalayam KR (2022) Iot-based smart healthcare video surveillance system using edge computing. J Ambient Intell Humaniz Comput 13(6):3195–3207
57. Li S, Song X, Xu S, Qi H, Xue Y (2023) Dilated spatial-temporal convolutional auto-encoders for human fall detection in surveillance videos. ICT Express 9(4):734–740
58. Khan SS, Mishra PK, Javed N, Ye B, Newman K, Mihailidis A, Iaboni A (2022) Unsupervised deep learning to detect agitation from videos in people with dementia. IEEE Access 10:10349–10358
59. Hao Y, Tang Z, Alzahrani B, Alotaibi R, Alharthi R, Zhao M, Mahmood A (2021) An end-to-end human abnormal behavior recognition framework for crowds with mentally disordered individuals. IEEE J Biomed Health Inform 26(8):3618–3625
60. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
61. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: Artificial neural networks and machine learning–ICANN 2018: 27th international conference on artificial neural networks Rhodes Greece October 4-7 2018 Proceedings Part III 27, Springer, pp 270–279
62. Vrbančič G, Podgorelec V (2020) Transfer learning with adaptive fine-tuning. IEEE Access 8:196197–196211
63. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
64. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
65. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
66. Carreira J, Zisserman A (2017) Quo vadis action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308
67. Howard AG (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
68. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
69. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T (2017) Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2462–2470
70. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90
71. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
72. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European conference Amsterdam The Netherlands October 11–14 2016 Proceedings Part I 14, Springer, pp 21–37
73. Feichtenhofer C (2020) X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 203–213

74. Teed Z, Deng J (2020) Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European conference Glasgow UK August 23–28 2020 Proceedings Part II 16, Springer, pp 402–419
75. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255
76. Lalit R, Purwar RK, Verma S, Jain A (2022) Crowd abnormality detection in video sequences using supervised convolutional neural network. Multimed Tools Appl 81(4):5259–5277
77. Gayal BS, Patil SR (2023) Detection and localization of anomalies in video surveillance using novel optimization based deep convolutional neural network. Multimed Tools Appl 82(19):28895–28915
78. Mehmood A (2021) Efficient anomaly detection in crowd videos using pre-trained 2d convolutional neural networks. IEEE Access 9:138283–138295
79. Zahid Y, Tahir MA, Durrani MN (2020) Ensemble learning using bagging and inception-v3 for anomaly detection in surveillance videos. In: 2020 IEEE international conference on image processing (ICIP), IEEE, pp 588–592
80. Shao W, Xiao R, Rajapaksha P, Wang M, Crespi N, Luo Z, Minerva R (2023) Video anomaly detection with ntcn-ml: A novel tcn for multi-instance learning. Pattern Recogn. 143:109765
81. Luo W, Liu W, Lian D, Tang J, Duan L, Peng X, Gao S (2019) Video anomaly detection with sparse coding inspired deep neural networks. IEEE Trans Pattern Anal Mach Intell 43(3):1070–1084
82. Zhang D, Huang C, Liu C, Xu Y (2022) Weakly supervised video anomaly detection via transformer-enabled temporal relation learning. IEEE Signal Process Lett 29:1197–1201
83. Ali MM (2023) Real-time video anomaly detection for smart surveillance. IET Image Proc 17(5):1375–1388
84. Bi Y, Li D, Luo Y (2022) Combining keyframes and image classification for violent behavior recognition. Appl Sci 12(16):8014
85. Ul Amin S, Ullah M, Sajjad M, Cheikh FA, Hijji M, Hijji A, Muhammad K (2022) Eadn: An efficient deep learning model for anomaly detection in videos. Math 10(9):1555
86. Vosta S, Yow KC (2024) Kiannet: A violence detection model using an attention-based cnn-lstm structure. IEEE Access
87. Butt UM, Letchmunan S, Hassan FH, Zia S, Baqir A (2020) Detecting video surveillance using vgg19 convolutional neural networks. Int J Adv Comput Sci Appl 11(2) (2020)
88. Zhang Z, Zhong Sh, Liu Y (2021) Video abnormal event detection via context cueing generative adversarial network. In: 2021 IEEE international conference on multimedia and expo (ICME), IEEE, pp 1–6
89. Zhong Y, Chen X, Hu Y, Tang P, Ren F (2022) Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. IEEE Trans. Circuits Syst. Video Technol 32(12):8285–8296
90. Liu HC, Chuah JH, Khairuddin ASM, Zhao XM, Wang XD (2023) Campus abnormal behavior recognition with temporal segment transformers (march 2023). IEEE Access
91. Lee S, Kim HG, Ro YM (2019) Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. IEEE Trans Image Process 29:2395–2408
92. Sabokrou M, Fayyaz M, Fathy M, Moayed Z, Klette R (2018) Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Comput Vis Image Underst 172:88–97
93. Hu ZP, Zhang L, Li SF, Sun DG (2020) Parallel spatial-temporal convolutional neural networks for anomaly detection and location in crowded scenes. Journal of Visual Communication and Image Representation. 67:102765
94. Sabokrou M, Fathy M, Hoseini M (2016) Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. Electron Lett 52(13):1122–1124
95. Li N, Chang F (2019) Video anomaly detection and localization via multivariate gaussian fully convolution adversarial autoencoder. Neurocomputing 369:92–105
96. Kotkar VA, Sucharita V (2023) Fast anomaly detection in video surveillance system using robust spatiotemporal and deep learning methods. Multimed Tools Appl 82(22):34259–34286
97. Ko KE, Sim KB (2018) Deep convolutional framework for abnormal behavior detection in a smart surveillance system. Eng Appl Artif Intell 67:226–234
98. Hamdi S, Bouindour S, Snoussi H, Wang T, Abid M (2021) End-to-end deep one-class learning for anomaly detection in uav video stream. J Imaging 7(5):90
99. Wu P, Liu J, Li M, Sun Y, Shen F (2020) Fast sparse coding networks for anomaly detection in videos. Pattern Recogn 107:107515
100. Yang F, Yu Z, Chen L, Gu J, Li Q, Guo B (2021) Human-machine cooperative video anomaly detection. Proceedings of the ACM on human-computer interaction 4(CSCW3):1–18
101. Rezaei F, Yazdi M (2021) Real-time crowd behavior recognition in surveillance videos based on deep learning methods. J Real-Time Image Proc 18(5):1669–1679
102. Jiang L, Zou B, Liu S, Yang W, Wang M, Huang E (2023) Recognition of abnormal human behavior in dual-channel convolutional 3d construction site based on deep learning. Neural Comput Appl 35(12):8733–8745
103. Rendón-Segador FJ, Álvarez-García JA, Enríquez F, Deniz O (2021) Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence. Electron 10(13):1601
104. Xu D, Yan Y, Ricci E, Sebe N (2017) Detecting anomalous events in videos by learning deep representations of appearance and motion. Comput Vis Image Underst 156:117–127
105. Doshi K, Yilmaz Y (2021) Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. Pattern Recognit 114:107865
106. Li J, Huang Q, Du Y, Zhen X, Chen S, Shao L (2021) Variational abnormal behavior detection with motion consistency. IEEE Trans Image Process 31:275–286
107. Hu X, Lian J, Zhang D, Gao X, Jiang L, Chen W (2022) Video anomaly detection based on 3d convolutional auto-encoder. SIViP 16(7):1885–1893
108. Jiang Z, Song G, Qian Y, Wang Y (2022) A deep learning framework for detecting and localizing abnormal pedestrian behaviors at grade crossings. Neural Comput Appl 34(24):22099–22113
109. Zeng X, Jiang Y, Ding W, Li H, Hao Y, Qiu Z (2021) A hierarchical spatio-temporal graph convolutional neural network for anomaly detection in videos. IEEE Trans Circuits Syst Video Technol 33(1):200–212
110. Liu C, Fu R, Li Y, Gao Y, Shi L, Li W (2021) A self-attention augmented graph convolutional clustering networks for skeleton-based video anomaly behavior detection. Appl Sci 12(1):4
111. Yang Y, Fu Z, Naqvi SM (2023) Abnormal event detection for video surveillance using an enhanced two-stream fusion method. Neurocomputing 553:126561
112. Song G, Qian Y, Wang Y (2023) Analysis of abnormal pedestrian behaviors at grade crossings based on semi-supervised generative adversarial networks. Appl Intell 53(19):21676–21691
113. Chen X, Kan S, Zhang F, Cen Y, Zhang L, Zhang D (2023) Multiscale spatial temporal attention graph convolution network for skeleton-based anomaly behavior detection. J Vis Commun Image Represent 90:103707

114. Al-Dhamari A, Sudirman R, Mahmood NH (2020) Transfer deep learning along with binary support vector machine for abnormal behavior detection. IEEE Access 8:61085–61095
115. Kokila MS, Christopher VB, Sajan RI, Akhila T, Kavitha MJ (2023) Efficient abnormality detection using patch-based 3d convolution with recurrent model. Mach Vis Appl 34(4):54
116. Kalshetty R, Parveen A (2023) Abnormal event detection model using an improved resnet101 in context aware surveillance system. Cogn Comput Syst 5(2):153–167
117. Chaurasia RK, Jaiswal UC (2023) Spatio-temporal based video anomaly detection using deep neural networks. Int J Inf Technol 15(3):1569–1581
118. Wu P, Liu J, Shen F (2019) A deep one-class neural network for anomalous event detection in complex scenes. IEEE Trans Neural Netw Learn Syst 31(7):2609–2622
119. Xia L, Li Z (2021) A new method of abnormal behavior detection using lstm network with temporal attention mechanism. J Supercomput 77(4):3223–3241
120. Georgescu MI, Barbalau A, Ionescu RT, Khan FS, Popescu M, Shah M (2021) Anomaly detection in video via self-supervised and multi-task learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12742–12752
121. Yang Z, Liu J, Wu P (2021) Bidirectional retrospective generation adversarial network for anomaly detection in videos. IEEE Access 9:107842–107857
122. Sernani P, Falcionelli N, Tomassini S, Contardo P, Dragoni AF (2021) Deep learning for automatic violence detection: Tests on the airtlab dataset. IEEE Access 9:160580–160595
123. Vu H, Nguyen TD, Le T, Luo W, Phung D (2019) Robust anomaly detection in videos using multilevel representations. Proceedings of the AAAI conference on artificial intelligence 33:5216–5223
124. Cho M, Kim T, Kim WJ, Cho S, Lee S (2022) Unsupervised video anomaly detection via normalizing flows with implicit latent features. Pattern Recogn 129:108703
125. Gallo G, Di Rienzo F, Garzelli F, Ducange P, Vallati C (2022) A smart system for personal protective equipment detection in industrial environments based on deep learning at the edge. IEEE Access 10:110862–110878
126. Petrocchi S, Giorgi G, Cimino MG (2021) A real-time deep learning approach for real-world video anomaly detection. In: Proceedings of the 16th international conference on availability reliability and security, pp 1–9
127. Kumar M, Biswas M (2023) Abnormal human activity detection by convolutional recurrent neural network using fuzzy logic. Multimed Tools Appl 1–17
128. Duman E, Erdem OA (2019) Anomaly detection in videos using optical flow and convolutional autoencoder. IEEE Access 7:183914–183923
129. Asad M, Yang J, Tu E, Chen L, He X (2021) Anomaly3d: Video anomaly detection based on 3d-normality clusters. J Vis Commun Image Represent 75:103047
130. Yan M, Meng J, Zhou C, Tu Z, Tan YP, Yuan J (2020) Detecting spatiotemporal irregularities in videos via a 3d convolutional autoencoder. J Vis Commun Image Represent 67:102747
131. Chang CW, Chang CY, Lin YY (2022) A hybrid cnn and lstm-based deep learning model for abnormal behavior detection. Multimed Tools Appl 81(9):11825–11843
132. Henrio J, Nakashima T (2018) Anomaly detection in videos recorded by drones in a surveillance context. In: 2018 IEEE international conference on systems man and cybernetics (SMC), IEEE, pp 2503–2508
133. Zhou JT, Du J, Zhu H, Peng X, Liu Y, Goh RSM (2019) Anomalynet: An anomaly detection network for video surveillance. IEEE Trans Inf Forensics Secur 14(10):2537–2550
134. Dong F, Zhang Y, Nie X (2020) Dual discriminator generative adversarial network for video anomaly detection. IEEE Access 8:88170–88176
135. Wang X, Che Z, Jiang B, Xiao N, Yang K, Tang J, Ye J, Wang J, Qi Q (2021) Robust unsupervised video anomaly detection by multipath frame prediction. IEEE Trans Neural Netw Learn Syst 33(6):2301–2312
136. Ding L, Fang W, Luo H, Love PE, Zhong B, Ouyang X (2018) A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. Autom Constr 86:118–124
137. Shin W, Bu SJ, Cho SB (2020) 3d-convolutional neural network with generative adversarial network and autoencoder for robust anomaly detection in video surveillance. Int J Neural Syst 30(06):2050034
138. Lei J, Sun W, Fang Y, Ye N, Yang S, Wu J (2024) A model for detecting abnormal elevator passenger behavior based on video classification. Electron 13(13):2472
139. Zhou K, Hui B, Wang J, Wang C, Wu T (2021) A study on attention-based lstm for abnormal behavior recognition with variable pooling. Image Vis Comput 108:104120
140. Maqsood R, Bajwa UI, Saleem G, Raza RH, Anwar MW (2021) Anomaly recognition from surveillance videos using 3d convolution neural network. Multimed Tools Appl 80(12):18693–18716
141. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst 28
142. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
143. Fang HS, Xie S, Tai YW, Lu C (2017) Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision, pp 2334–2343
144. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
145. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5693–5703
146. Wojke N, Bewley A, Paulus D (2017) Simple online and real-time tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP), IEEE, pp 3645–3649
147. Xiu Y, Li J, Wang H, Fang Y, Lu C (2018) Pose flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977
148. Welch G, Bishop G et al (1995) An introduction to the kalman filter
149. Pang G, Yan C, Shen C, Hengel AVD, Bai X (2020) Self-trained deep ordinal regression for end-to-end video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12173–12182
150. Chen D, Yue L, Chang X, Xu M, Jia T (2021) Nm-gan: Noise-modulated generative adversarial network for video anomaly detection. Pattern Recogn 116:107969
151. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference Munich Germany October 5-9 2015 Proceedings Part III 18, Springer, pp 234–241
152. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497

153. Ullah FUM, Ullah A, Muhammad K, Haq IU, Baik SW (2019) Violence detection using spatiotemporal features with 3d convolutional neural network. Sensors 19(11):2472
154. Li B, Leroux S, Simoens P (2021) Decoupled appearance and motion learning for efficient anomaly detection in surveillance video. Comput Vis Image Underst 210:103249
155. Hao Y, Li J, Wang N, Wang X, Gao X (2022) Spatiotemporal consistency-enhanced network for video anomaly detection. Pattern Recogn 121:108232
156. Zhou S, Shen W, Zeng D, Fang M, Wei Y, Zhang Z (2016) Spatial-temporal convolutional neural networks for anomaly detection and localization in crowded scenes. Signal Processing: Image Commun 47:358–368
157. Doshi K, Yilmaz Y (2020) Continual learning for anomaly detection in surveillance videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 254–255
158. Feng Q, Gao C, Wang L, Zhao Y, Song T, Li Q (2020) Spatio-temporal fall event detection in complex scenes using attention guided lstm. Pattern Recogn Lett 130:242–249
159. Ullah W, Ullah A, Haq IU, Muhammad K, Sajjad M, Baik SW (2020) Cnn features with bi-directional lstm for real-time anomaly detection in surveillance networks. Multimed Tools Appl 80:16979–16995
160. Vijeikis R, Raudonis V, Dervinis G (2022) Efficient violence detection in surveillance. Sensors 22(6):2216
161. Qasim M, Verdu E (2023) Video anomaly detection system using deep convolutional and recurrent models. Results Eng 18:101026
162. Yang B, Cao J, Wang N, Liu X (2018) Anomalous behaviors detection in moving crowds based on a weighted convolutional autoencoder-long short-term memory network. IEEE Trans Cogn Develop Syst 11(4):473–482
163. Nawaratne R, Alahakoon D, De Silva D, Yu X (2019) Spatiotemporal anomaly detection using deep learning for real-time video surveillance. IEEE Trans Industr Inf 16(1):393–402
164. Li T, Chen X, Zhu F, Zhang Z, Yan H (2021) Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection. Neurocomputing 439:256–270
165. Pawar K, Attar V (2022) Deep learning model based on cascaded autoencoders and one-class learning for detection and localization of anomalies from surveillance videos. IET Biometrics 11(4):289–303
166. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 27
167. Cheng K, Liu Y, Zeng X (2023) Learning graph enhanced spatial-temporal coherence for video anomaly detection. IEEE Signal Process Lett 30:314–318
168. Aslam N, Kolekar MH (2024) Demaae: deep multiplicative attention-based autoencoder for identification of peculiarities in video sequences. Vis Comput 40(3):1729–1743
169. Luo W, Liu W, Gao S (2021) Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. Neurocomputing 444:332–337
170. Yu B, Yin H, Zhu Z (2017) Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv preprint arXiv:1709.04875
171. Yang Y, Fu Z, Naqvi SM (2022) A two-stream information fusion approach to abnormal event detection in video. In: ICASSP 2022-2022 IEEE international conference on acoustics speech and signal processing (ICASSP), IEEE, pp 5787–5791
172. Ravanbakhsh M, Nabi M, Sangineto E, Marcenaro L, Regazzoni C, Sebe N (2017) Abnormal event detection in videos using generative adversarial nets. In: 2017 IEEE international conference on image processing (ICIP), IEEE, pp 1577–1581
173. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6536–6545
174. Ravanbakhsh M, Sangineto E, Nabi M, Sebe N (2019) Training adversarial discriminators for cross-channel abnormal event detection in crowds. In: 2019 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 1896–1904
175. Ehsan TZ, Nahvi M, Mohtavipour SM (2024) An accurate violence detection framework using unsupervised spatial-temporal action translation network. Vis Comput 40(3):1515–1535
176. Farnebäck G (2003) Two-frame motion estimation based on polynomial expansion. In: Image Analysis: 13th scandinavian conference SCIA 2003 Halmstad Sweden June 29–July 2 2003 Proceedings 13, Springer, pp 363–370
177. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
178. Li S, Liu F, Jiao L (2022) Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. Proceedings of the AAAI conference on artificial intelligence 36:1395–1403
179. Rendón-Segador FJ, Álvarez-García JA, Salazar-González JL, Tommasi T (2023) Crimenet: neural structured learning using vision transformer for violence detection. Neural Netw 161:318–329
180. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H (2022) Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3202–3211
181. Joo HK, Vo K, Yamazaki K, Le N (2023) Clip-tsa: Clip-assisted temporal self-attention for weakly-supervised video anomaly detection. In: 2023 IEEE international conference on image processing (ICIP), IEEE, pp 3230–3234
182. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PMLR, pp 8748–8763
183. Zhu H, Wei P, Xu Z (2024) A spatio-temporal enhanced graph-transformer autoencoder embedded pose for anomaly detection. IET Comput Vision 18(3):405–419
184. Wu P, Zhou X, Pang G, Zhou L, Yan Q, Wang P, Zhang Y (2024) Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. Proceedings of the AAAI conference on artificial intelligence 38:6074–6082
185. Zanella L, Menapace W, Mancini M, Wang Y, Ricci E (2024) Harnessing large language models for training-free video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 18527–18536
186. Li J, Li D, Savarese S, Hoi S (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning, PMLR, pp 19730–19742
187. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971
188. Tan M, Le Q (2021) Efficientnetv2: Smaller models and faster training. In: International conference on machine learning, PMLR, pp 10096–10106
189. El Kafhali S, El Mir I, Hanini M (2022) Security threats defense mechanisms challenges and future directions in cloud computing. Arch Comput Methods Eng 29(1):223–246
190. Khan WZ, Ahmed E, Hakak S, Yaqoob I, Ahmed A (2019) Edge computing: A survey. Futur Gener Comput Syst 97:219–235

191. Gu J, Feng J, Xu H, Zhou T (2022) Research on terminal-side computing force network based on massive terminals. Electron 11(13):2108
192. Cheng H, Liu X, Wang H, Fang Y, Wang M, Zhao X (2020) Securead: A secure video anomaly detection framework on convolutional neural network in edge computing environment. IEEE Trans Cloud Comput 10(2):1413–1427
193. Chang WJ, Hsu CH, Chen LB (2021) A pose estimation-based fall detection methodology using artificial intelligence edge computing. IEEE Access 9:129965–129976
194. Osokin D (2018) Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. arXiv preprint arXiv:1811.12004
195. Mehmood A (2021) Lightanomalynet: a lightweight framework for efficient abnormal behavior detection. Sensors 21(24):8501
196. Kim JH, Won CS (2020) Action recognition in videos using pre-trained 2d convolutional neural networks. IEEE Access 8:60179–60188
197. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934
198. Nagel M, Fournarakis M, Amjad RA, Bondarenko Y, Van Baalen M, Blankevoort T (2021) A white paper on neural network quantization. arXiv preprint arXiv:2106.08295
199. Ullah W, Ullah A, Hussain T, Muhammad K, Heidari AA, Del Ser J, Baik SW, De Albuquerque VHC (2022) Artificial intelligence of things-assisted two-stream neural network for anomaly detection in surveillance big video data. Futur Gener Comput Syst 129:286–297
200. Lim WYB, Luong NC, Hoang DT, Jiao Y, Liang YC, Yang Q, Niyato D, Miao C (2020) Federated learning in mobile edge networks: A comprehensive survey. IEEE Commun Surv Tutor 22(3):2031–2063
201. Rodrigues R, Bhargava N, Velmurugan R, Chaudhuri S (2020) Multi-timescale trajectory prediction for abnormal human activity detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2626–2634
202. Ramachandra B, Jones M (2020) Street scene: A new dataset and evaluation protocol for video anomaly detection. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2569–2578
203. Adhikari K, Bouchachia H, Nait-Charif H (2017) Activity recognition for indoor fall detection using convolutional neural network. In: 2017 15th IAPR international conference on machine vision applications (MVA), IEEE, pp 81–84
204. Auvinet E, Rougier C, Meunier J, St-Arnaud A, Rousseau J (2010) Multiple cameras fall dataset. DIRO-Université de Montréal Tech Rep 1350:24
205. Kwolek B, Kepski M (2014) Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput Methods Prog Biomed 117(3):489–501
206. Charfi I, Miteran J, Dubois J, Atri M, Tourki R (2012) Definition and performance evaluation of a robust svm based fall detection solution. In: 2012 18th international conference on signal image technology and internet based systems, IEEE, pp 218–224
207. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 1975–1981. https://doi.org/10.1109/CVPR.2010.5539872
208. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision, pp 2720–2727
209. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked rnn framework. In: Proceedings of the IEEE international conference on computer vision, pp 341–349
210. Wu P, Liu J, Shi Y, Sun Y, Shao F, Wu Z, Yang Z (2020) Not only look but also listen: Learning multimodal violence detection under weak supervision. In: Computer cision–ECCV 2020: 16th European conference Glasgow UK August 23–28 2020 Proceedings Part XXX 16, Springer, pp 322–339
211. Bonetto M, Korshunov P, Ramponi G, Ebrahimi T (2015) Privacy in mini-drone based video surveillance. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 4. IEEE, pp 1–6
212. Bermejo Nievas E, Deniz Suarez O, Bueno García G, Sukthankar R (2011) Violence detection in video using computer vision techniques. In: Computer analysis of images and patterns: 14th international conference CAIP 2011 Seville Spain August 29-31 2011 Proceedings Part II 14, Springer, pp 332–339
213. Cheng M, Cai K, Li M (2021) Rwf-2000: an open large scale video database for violence detection. In: 2020 25th international conference on pattern recognition (ICPR), IEEE, pp 4183–4190
214. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 935–942
215. Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans Pattern Anal Mach Intell 30(3):555–560
216. Hassner T, Itcher Y, Kliper-Gross O (2012) Violent flows: Real-time detection of violent crowd behavior. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE, pp 1–6
217. Ferryman J, Shahrokni A (2009) Pets2009: Dataset and challenge. In: 2009 12th IEEE international workshop on performance evaluation of tracking and surveillance, IEEE, pp 1–6
218. Singh D, Mohan CK (2018) Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. IEEE Trans Intell Transp Syst 20(3):879–887
219. Blunsden S, Fisher R (2010) The behave video dataset: Ground truthed video for multi-person behavior classification. Ann BMVA 4(1–12):4
220. Ryoo MS, Aggarwal JK (2009) Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: 2009 IEEE 12th international conference on computer vision, IEEE, pp 1593–1600
221. Sabokrou M, Fayyaz M, Fathy M, Klette R (2017) Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans Image Process 26(4):1992–2004
222. Nguyen TN, Meunier J (2019) Anomaly detection in video sequence with appearance-motion correspondence. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1273–1283
223. Kamoona AM, Gostar AK, Bab-Hadiashar A, Hoseinnezhad R (2023) Multiple instance-based video anomaly detection using deep temporal encoding-decoding. Expert Syst Appl 214:119079
224. Zanella L, Liberatori B, Menapace W, Poiesi F, Wang Y, Ricci E (2024) Delving into clip latent space for video anomaly recognition. Comput Vis Image Underst 249:104163
225. Girdhar R, El-Nouby A, Liu Z, Singh M, Alwala KV, Joulin A, Misra I (2023) Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 15180–15190

226. Zhu Y, Bao W, Yu Q (2022) Towards open set video anomaly detection. In: European conference on computer vision, Springer, pp 395–412
227. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas DDL, Bressand F, Lengyel G, Lample G, Saulnier L et al (2023) Mistral 7b. arXiv preprint arXiv:2310.06825
228. OpenAI (2023) GPT-4V(ision) system card. https://openai.com/index/gpt-4v-system-card/
229. Ye M, Liu W, He P (2024) Vera: Explainable video anomaly detection via verbalized learning of vision-language models. arXiv preprint arXiv:2412.01095
230. Ding X, Wang L (2024) Quo vadis anomaly detection? llms and vlms in the spotlight. arXiv preprint arXiv:2412.18298