Multimedia Tools and Applications (2024) 83:59617–59674
https://doi.org/10.1007/s11042-023-17889-z

A comprehensive review of datasets for detection and localization of video anomalies: a step towards data-centric artificial intelligence-based video anomaly detection

Rashmiranjan Nayak[1] · Umesh Chandra Pati[1] · Santos Kumar Das[1]

Received: 3 September 2022 / Revised: 3 October 2023 / Accepted: 13 December 2023 /
Published online: 29 December 2023

Abstract

Video anomaly detection and localization is one of the key components of the intelligent video surveillance system. Video anomaly detection refers to the process of spatiotemporal localization of the abnormal or anomalous pattern present in the video. The performance of the deep learning-based video anomaly detector depends on the quality and quantity of the video anomaly datasets used for training. However, there is a scarcity of effective video anomaly datasets due to inherent natures such as rareness, context-dependency, and equivocal nature. Further, state-of-the-art lacks a review that presents a comprehensive study of video anomaly datasets, including issues associated with the existing datasets, comparative analysis of the available datasets, potential solutions using both model-centric and data-centric approaches. Hence, a comprehensive review of the publicly available video anomaly datasets for video anomaly detection and localization is presented in this article. Further, a comparative study of the existing video anomaly datasets at qualitative and quantitative levels is presented to decide the right strategies for the desired application. Subsequently, model-centric and data-centric approaches required to solve various problems associated with the video anomaly datasets are presented. Finally, current research trends, research challenges, potential applications, and future research directions are outlined.

Keywords Deep learning · Data-centric approach · Model-centric approach · Video anomaly detection · Video anomaly localization · Video anomaly datasets

✉ Rashmiranjan Nayak
rashmiranjan.et@gmail.com

Umesh Chandra Pati
ucpati@nitrkl.ac.in

Santos Kumar Das
dassk@nitrkl.ac.in

[1] Department of Electronics and Communication Engineering, National Institute of Technology Rourkela, Sector-1, Rourkela 769008, Odisha, India

1 Introduction

Intelligent Video Surveillance System (IVSS) is a technology on demand to ensure the safety and security of the lives as well as assets in public places such as market places, shopping malls, hospitals, banks, streets, educational institutions, city administrative offices, smart cities, etc. Generally, most of the videos recorded by the surveillance cameras corresponding to the normal events are not important [1]. However, the events or entities which are abnormal (or anomalous) in nature are of greater importance for intelligent video analytic [2, 3]. Hence, automatic detection of anomalous events or entities (video anomalies) using surveillance video is one of the major tasks of the IVSS. The video anomalies are defined as the irregular Spatio-temporal patterns available in the video that significantly deviates from the normal trained patterns. The rapid growth in Artificial Intelligence (AI) technologies such as Machine Learning (ML) and Deep Learning (DL) have enabled the IVSS to detect and localize the video anomalies efficiently in a complete automatic framework. The availability of large video datasets, better computational facilities, and inventions of enhanced algorithms are the three important driving force in achieving the automatic detection as well as localization of the video anomalies. The detection and localization of the video anomalies may be defined as the process of detecting as well as localizing the video anomalies in the Spatio-temporal dimensions. In other words, the complete process involves two important steps such as Video Anomaly Detection (VAD) and Video Anomaly Localization (VAL). The video anomaly detection focuses on finding whether the given video frame exhibits an anomaly or not. Subsequently, the video anomaly localization focuses on the localization of anomalies by determining the actual location of the anomalies in the given video frame [4]. Moreover, due to the advancement of DL methods capable of performing detection and localization of the video anomalies in a single end-to-end pipeline, sometimes only VAD signifies both detection and localization of the video anomalies. Further, the detection and localization of the video anomalies are always decided based on the selected evaluation criteria. Then, the video anomaly detector is evaluated in terms of the various performance metrics for the quantitative as well as qualitative analysis.

In this section, we first discuss the motivations and scope of this survey. Subsequently, we present a brief discussion concerning the related surveys and contributions made in this article.

1.1 Motivations

Deep learning is part of machine learning where data-driven approaches are used instead of feature engineering. Recently, deep learning techniques are widely used for the detection of video anomalies to provide high accuracy and dataset generalization as compared to ML techniques based on feature engineering [4]. Generally, the performance of the deep-learning models is primarily dependent on the quantity as well as the quality of the datasets that are used in modeling [5, 6]. Hence, the performance of the deep learning-based models that are developed for the video anomaly detection and localization is significantly dependent on the quantity as well as the quality of the video anomaly datasets. However, there are few essential inherent issues concerning the video anomaly datasets which deters the development, performance, and applicability of the deep-learning-based video anomaly detectors. These are listed as follows.

- The video anomaly detection and localization help in reducing the human effort by providing a coarse level of video understanding by filtering out the anomalous patterns from the normal patterns promptly [7]. Subsequently, a detailed video analysis involving object detection, activity classification, and person recognition can be performed only for the anomalous video segment instead of the whole video. Hence, automatic video anomaly detection and localization are helpful for humankind in multidimensional applications by reducing the manual effort [8].

- Though many datasets are available, there is a lack of a comprehensive analysis of these datasets highlighting the individual pros and cons of corresponding applications [8].
- The equivocal nature of the video anomalies prevents the perfect labeling in the Spatio-temporal dimensions and results in a dearth of anomalous ground truth data (or insufficient labeled data). Hence, the development of the end-to-end trainable deep learning-based models for the detection and localization of anomalies are restricted [9–11].
- Video anomalies are context-dependent [12, 13] and hence, the development of a universal video anomaly detector is not a feasible option. Therefore, the development of the video anomaly detector is constrained by the application domain, environmental conditions, and the scope of training data.

The points, as mentioned earlier, are the few prominent driving forces for preparing the article.

1.2 Scope of this survey

The following factors define the scope of the survey.

- There are a variety of datasets available for anomaly detection corresponding to the various domains. However, this survey will be restricted to only datasets of video anomaly detection and localization for providing a comprehensive analysis.
- Broadly, the datasets of the video anomalies can be captured either with the help of a stationary (or fixed) surveillance camera and dynamic (or moving) camera. However, this survey will be mostly restricted to datasets corresponding to the stationary surveillance camera.
- The survey will cover most of the prominent datasets available for both the single scene and multiple scene problem formulation.
- The preprocessing techniques which are suitable only for deep learning methods will be discussed.

1.3 Related surveys

There are few existing surveys [2, 4, 12, 14–20] in the area of video anomaly detection and localization. However, as per the best of our knowledge, there is only one survey [8] dedicated to the datasets used in the video anomaly detection. The tremendous progress in deep learning-based video anomaly detection in the last five years has been accompanied by the creation of good quality new video anomaly datasets. However, the previous survey [8] has not discussed the issues in the existing datasets, prominent preprocessing techniques, open research challenges, and potential application domains. The present review addresses all the points mentioned above while including all the recent datasets.

1.4 Contributions

The present review is initially built on the previous survey related to the video datasets for anomaly detection [8]. Subsequently, the article presents a detailed and structured analysis of datasets reported in the state-of-the-art for the detection and localization of the video anomalies. The significant contributions of the research work can be summarized as follows.

- A comprehensive review of the bench-marked video datasets that are reported in the state-of-the-art research for video anomaly detection is presented.
- A comparative analysis of the datasets used for the detection and localization of the video anomalies has been put forth. This will help the research community to identify the appropriate dataset for the desired application.
- A summary of problem formulation for the video anomaly detection depending on the training video anomaly datasets using either Single Scene Formulation or Multiple Scene Formulation is presented.
- A brief analysis of the AI approaches, i.e., data-centric approaches in terms of various data preprocessing techniques and model-centric approaches in terms of various model enhancement techniques that enhance the model performance for the existing video anomaly datasets, are presented.
- Emerging trends in the detection and localization of the video anomalies and corresponding alignment with the video anomaly datasets are briefly outlined.
- The prominent problems, open research challenges and future directions of the video anomaly detection and localization associated with the existing video anomaly datasets have been outlined.

1.5 Organization

The rest of the article is structured as follows. Preliminaries required for the video anomaly detection and localization are briefly presented in Section 2. Problem formulation for video anomaly detection depending on the datasets used for training is summarized in Section 3. Classification and comparative analysis of datasets for the video anomaly detection are presented in Sections 4 and 5, respectively. Subsequently, the issues with existing datasets are explained in Section 6, followed by the various data explorations and preparation techniques required for understanding as well as transforming the data in Section 7. AI approaches in terms of model-centric and data-centric approaches to develop efficient DL models for the detection as well as localization of the video anomalies by solving issues related to the video anomaly datasets are discussed in Section 8. The emerging trends in the detection and localization of the video anomalies as per the availability of the video anomaly datasets are presented in Section 9. The potential applications of video anomaly detection are summarized in Section 10. The open research challenges and future directions specifically due to the problems associated with the video anomaly datasets are outlined in Section 11. Finally, a brief conclusion is presented in Section 12.

2 Preliminaries for the detection and localization of video anomalies

This Section presents essential preliminaries required to understand the detection and localization of video anomalies properly.

2.1 Unique characteristics of video anomalies

Anomalies are one of the subsets of outliers [21]. Outliers are the deviants comprised of noise and anomalies. However, noise is treated as the uninteresting outlier, whereas anomalies are considered as sufficiently interesting outliers. Video anomalies are equivocal (whose interpretation is highly contextually dependent) [22], novel [23], unknown, rare [11], unexpected [24], atypical [25], abnormal, and out-of-the-dictionary [15, 26–28] in nature.

2.2 Complexity of the surveillance scenarios

The complexity of the surroundings and the types of anomalies are two of the most important considerations in video anomaly detection and localization. Furthermore, the density of moving objects contributes to the complexity of the scene. Consequently, the density of the moving targets allows us to categorize the environment into three distinct types as follows.

2.2.1 Sparsely crowded environment

When objects are loosely distributed, i.e., around 10 square feet of area is available per person, then such an environment is known as a sparsely crowded environment [29]. Video anomalies involving single objects, such as loitering, intrusion, etc., and involving interactions of two persons, such as fighting, are typical examples of the sparsely crowded environment.

2.2.2 Moderately crowded environment

When objects are relatively more crowded as compared to that moderately crowded environment, i.e., around 4.5 square feet of area is available per person, then such an environment is known as a moderately crowded environment [29]. Video anomalies involving group activities, such as riots, violence, etc., are typical examples of the sparsely crowded environment.

2.2.3 Densely crowded environment

When objects are highly populated or densely crowded, i.e., around 2.5 square feet of area is available per person, then such an environment is known as a densely crowded environment [29]. Video anomalies involving highly packed crowd activities, such as stampedes, sudden crowd dispersion caused by an explosion, etc., are typical examples of a densely crowded environment.

2.3 Types of video anomalies

Video anomalies may be local or global based on the level of occurrences [4, 19, 22, 30–32]. Further, video anomalies may be point or interaction anomalies based on the number of objects involved [4, 19, 22]. However, video anomalies are spatiotemporal anomalies that can be best described by contextual or conditional anomalies [4, 19]. Contextual anomalies correspond to the data points having significant deviations causing anomalies with respect to a specific context defined by the contextual and behavioral features [33]. Usually, time and location are examples of contextual factors, whereas characteristics that reflect typical conduct are examples of behavioral features. Most of the video anomalies, such as stampedes, riots, violence, fighting, etc., can be suitably described by contextual anomalies with the help of spatiotemporal (combination of appearance and motion) features [22, 34].

2.4 Learning approaches

Based on the level of human intervention and utilization of labeled data during training, video anomaly detectors can be trained using any one of the following techniques.

2.4.1 Supervised video anomaly detection and localization

Supervised video anomaly detection and localization methods involve training a binary classifier using the associated labels of normal and anomaly frames as well as pixels using a balanced training dataset comprised of clearly defined events. Unfortunately, it is not practically feasible to perfectly define video anomalies due to their ambiguous nature, evolutionary quality, inherent data-imbalance problem, and probability of high variance within anomalies [19, 22].

2.4.2 Unsupervised video anomaly detection and localization

Unsupervised video anomaly detection and localization methods involve the usage of unlabeled video data that rely on co-occurrence statistical concepts to identify suspicious events or objects [15]. Furthermore, unsupervised video anomaly detection and localization algorithms often need a large video dataset and substantial computing resources to be successful [22]. Because these two resources are so readily available, unsupervised video anomaly detection systems have consistently outperformed their supervised counterparts. While enormous amounts of weakly labeled normal data are accessible, labeled data for solely abnormal activity is scarce. However, the unsupervised video anomaly detection methods do not exploit the full potential of this poorly labeled normal data.

2.4.3 Semi-supervised video anomaly detection and localization

This sort of video anomaly detection and localization technique is increasingly popular since it combines the advantages of supervised and unsupervised methods. Due to the availability of normal videos or data devoid of abnormalities, unsupervised video anomaly identification methods are often handled as semi-supervised video anomaly detection approaches [12]. Deep-autoencoder based models have recently been trained with enough training data that solely includes normal events, resulting in models with the lowest possible reconstruction error for normal activities [19, 22]. As a result, the model detects and localizes the video abnormalities since it provides a significantly high reconstruction error for abnormal activity. However, by keeping a domain expert informed, the effectiveness of the semi-supervised video anomaly detection algorithms may be increased even further.

2.4.4 Active learning-based video anomaly detection and localization

The model is trained offline using typical training examples in the case of unsupervised or semi-supervised video anomaly detection and localization, and it is not updated as and when fresh data is received [35]. As a consequence, the resulting audiovisual representations are ineffectual. The use of active learning-based video anomaly detection, which keeps humans (or domain experts) in the loop for categorizing the perplexing choices or samples in the online framework, may be used to overcome these challenges. Thus, by incorporating suitable priors with the aid of a domain expert, active learning aids in decreasing the ambiguous character of anomaly [36]. A deep active learning approach has recently been put out for unsupervised deep learning-based anomaly detection models. Additionally, the Generative Adversarial Network (GAN) features are effectively used for the outlier identification process [22, 37]. Despite becoming more accurate in online applications, active learning-based video anomaly detection algorithms still need ongoing input from subject-matter experts.

2.5 Targeted applications

Desired accuracy and speed (processing time) of the models meant for the detection and localization of the video anomalies are decided by the targeted application. Broadly targeted applications can be classified as either online applications or offline applications.

2.5.1 Online applications

Online applications involve detecting and localizing video anomalies from live video streams. Here, the objective is to detect and localize the video anomalies with high speed (or the least frame processing time) and competitive accuracy. In other words, the current frame of the live video stream must be entirely processed before the arrival of the next frame by optimizing the frame processing time [38]. Models suitable for online (real-time) applications attain online performances such as high computational speed and less computation space by employing lightweight and robust features [19, 22].

2.5.2 Offline applications

Offline applications involve detecting and localizing video anomalies from offline or stored videos. Here, the objective is to detect and localize the video anomalies with high accuracy and minimum false alarms by optimizing the detection accuracy [38]. Models suitable for offline applications attain the highest accuracy at the expense of higher computational time by employing more complex and descriptive features [19, 22].

3 Problem formulation for video anomaly detection

Generally, depending on the available datasets, the problem formulation for the video anomaly detection can be qualitatively divided into two main categories as Single Scene Formulation (SSF) and Multiple scene Formulation (MSF) [1].

3.1 Video anomaly detection using single scene formulation

In the case of video anomaly detection using SSF, the model is trained using the training videos V_{train} or frame sequences consists F_{train} of only normal (non-anomalous) events from a single scene. Subsequently, the trained video anomaly detector is used to detect the video anomalies from the test videos or frame sequences from the same scene [1]. Here, video anomalies signify the spatiotemporally localized video segment that differs significantly from the trained patterns using the training videos. The exact quantification of the “significantly different" is very difficult to specify universally and is dependent on the target application. Mainly, this significant difference is caused due to the appearance (spatial) and temporal (motion) of the objects present in the video corresponding to the particular scene.

Mathematically, let us consider that the complete video dataset V_{SS} is collected from the same scenario. Hence, number of scenes available in the V_{SS} is one. In other words, total number of scenes being considered in the video anomaly detection using single scene formulation is one,i.e., $N_{SS} = 1$. Hence, complete video dataset V_{SS} and total number of video clips available in the V are presented in (1) and (2), respectively.

$$V_{SS} = V_{train}^{SS} \bigcup V_{test}^{SS} \tag{1}$$

$$N_{V}^{SS} = N_{train}^{SS} + N_{test}^{SS} \tag{2}$$

Further, the training video dataset V_{train}^{SS} may be comprised of multiple video clips as represented in (3). Here, $v_{train_j}^{SS}$ represents the j^{th} video clip of the training video set V_{train}^{SS} and N_{train}^{SS} is the number of video clips available in the V_{train}^{SS}.

$$V_{train}^{SS} = \bigcup_{j=1}^{N_{train}^{SS}} v_{train_j}^{SS} = \left\{ v_{train_1}^{SS}, v_{train_2}^{SS}, v_{train_3}^{SS}, \ldots, v_{train_{N_{train}^{SS}}}^{SS} \right\} \tag{3}$$

Each $v_{train_j}^{SS}$ comprises multiple individual frame $f_{train_k}^{SS}$ as expressed in (4). Here, M_{train}^{SS} is the total number of the frames available in each $v_{train_j}^{SS}$.

$$v_{train_j}^{SS} = \bigcup_{k=1}^{M_{train}^{SS}} f_{train_k}^{SS} = \left\{ f_{train_1}^{SS}, f_{train_2}^{SS}, f_{train_3}^{SS}, \ldots, f_{train_{M_{train}^{SS}}}^{SS} \right\} \tag{4}$$

Now, V_{train}^{SS} can be rewritten by combining (3) and (4) as expressed in (5) and (6).

$$V_{train}^{SS} = \bigcup_{j=1}^{N_{train}^{SS}} \left\{ \bigcup_{k=1}^{M_{train}^{SS}} f_{train_{jk}}^{SS} \right\} \tag{5}$$

$$\Rightarrow V_{train}^{SS} = \bigcup_{j=1}^{N_{train}^{SS}} \left\{ f_{train_{j1}}^{SS}, f_{train_{j2}}^{SS}, f_{train_{j3}}^{SS}, \ldots, f_{train_{jM_{train}^{SS}}}^{SS} \right\} \tag{6}$$

Similarly, the testing video dataset V_{test}^{SS} consists of multiple video clips as represented in (7). Here, $v_{test_j}^{SS}$ represents the j^{th} video clip of the testing video set V_{test}^{SS} and N_{test}^{SS} is the number of video clips available in the V_{test}^{SS}.

$$V_{test}^{SS} = \bigcup_{j=1}^{N_{test}^{SS}} v_{test_j}^{SS} = \left\{ v_{test_1}^{SS}, v_{test_2}^{SS}, v_{test_3}^{SS}, \ldots, v_{test_{N_{test}^{SS}}}^{SS} \right\} \tag{7}$$

Each $v^{SS}_{test_j}$ comprises multiple individual frame $f^{SS}_{test_k}$ as expressed in (8). Here, M^{SS}_{test} is the total number of the frames available in each $v^{SS}_{test_j}$.

$$v^{SS}_{test_j} = \bigcup_{k=1}^{M^{SS}_{test}} f^{SS}_{test_k} = \left\{ f^{SS}_{test_1}, f^{SS}_{test_2}, f^{SS}_{test_3}, \ldots, f^{SS}_{test_{M^{SS}_{test}}} \right\} \tag{8}$$

Now, V^{SS}_{test} can be rewritten by combining (7) and (8) as expressed in (9) and (10).

$$V^{SS}_{test} = \bigcup_{j=1}^{N^{SS}_{test}} \left\{ \bigcup_{k=1}^{M^{SS}_{test}} f^{SS}_{test_{jk}} \right\} \tag{9}$$

$$\Rightarrow V^{SS}_{test} = \bigcup_{j=1}^{N^{SS}_{test}} \left\{ f^{SS}_{test_{j1}}, f^{SS}_{test_{j2}}, f^{SS}_{test_{j3}}, \ldots, f^{SS}_{test_{jM^{SS}_{test}}} \right\} \tag{10}$$

Generally, video anomaly detection is treated as an unsupervised learning problem as there is no direct information about the anomaly classes (positive classes). However, direct information about the normal classes (negative classes) is available in most practical scenarios. Hence, the video anomaly detection problem can be further simplified and treated as a semi-supervised learning problem. The normal class distribution D_N can be estimated by using the training samples V_{train} consisting of only normal classes by modeling an automatic representation (Video Anomaly Detector) VAD_{SS} that minimizes the reconstruction cost. Mathematically, the objective function for the video anomaly detection can be formulated as represented in (11).

$$minimize \ \left\| f^{SS}_{train_{jk}} - VAD_{SS}\left(f^{SS}_{train_{jk}} \right) \right\|^2 \tag{11}$$

subject to constraints

$$1 \leqslant j \leqslant N^{SS}_{train} \tag{12}$$

$$1 \leqslant k \leqslant M^{SS}_{train} \tag{13}$$

Once the model is learned, the testing frame sequences $f^{SS}_{test_{jk}}$ are fed into the trained model VAD_{SS} to reconstruct the frames $\hat{f}^{SS}_{test_{jk}}$ as compressed in (14),

$$\hat{f}^{SS}_{test_{jk}} = VAD_{SS}\left(f^{SS}_{test_{jk}} \right) \tag{14}$$

where,

$$1 \leqslant j \leqslant N^{SS}_{test} \tag{15}$$

$$1 \leqslant k \leqslant M^{SS}_{test}. \tag{16}$$

The reconstruction error for a given pixel with an intensity I at a spatial location (x, y) in a particular testing frame $f^{SS}_{test_{jk}}$ at time instant t can be calculated by using the learned model VAD_{SS} by using (18) [39].

$$e^{SS}_{test_{jk}}(x, y, t) = \left\| I^{SS}_{test_{jk}}(x, y, t) - \hat{I}^{SS}_{test_{jk}}(x, y, t) \right\|^2 \tag{17}$$

$$= I^{SS}_{test_{jk}}(x, y, t) - VAD_{SS}\left(I^{SS}_{test_{jk}}(x, y, t) \right) \tag{18}$$

Here, $I_{test_{jk}}^{SS}(x, y, t)$ and $\hat{I}_{test_{jk}}^{SS}(x, y, t)$ are the pixel intensities at the spatiotemporal location (x, y, t) corresponding to the frames $f_{test_{jk}}^{SS}$ and $\hat{f}_{test_{jk}}^{SS}$, respectively. Further, the reconstruction error of the particular test frame at time t , $i.e.$, $e_{test_{jk}}^{reconstruct_{SS}}(t)$ can be calculated from the known pixel-level reconstruction errors $e_{test_{jk}}^{SS}(x, y, t)$ by using (19) [39].

$$e_{test_{jk}}^{reconstruct_{SS}}(t) = \sum_{x,y} e_{test_{jk}}^{SS}(x, y, t) \tag{19}$$

Subsequently, the anomaly score $S_{test_{jk}}^{ano_{SS}}(t)$ and the regularity score $S_{test_{jk}}^{reg_{SS}}(t)$ for the test frame $test_{jk}^{SS}(t)$ can be calculated by using (20) [40] and (21) [39]. Here, the values of both anomaly and regularity scores are lay in the range of 0 to 1.

$$S_{test_{jk}}^{ano_{SS}}(t) = \frac{e_{test_{jk}}^{reconstruct_{SS}}(t) - e_{test_{jk}}^{reconstruct_{SS_{min}}}(t)}{e_{test_{jk}}^{reconstruct_{SS_{max}}}(t)} \tag{20}$$

$$S_{test_{jk}}^{reg_{SS}}(t) = 1 - S_{test_{jk}}^{ano_{SS}}(t) \tag{21}$$

Finally, the individual test frames are considered anomalous or normal based on the associated anomaly score by using the condition mentioned in the (22).

$$S_{test_{jk}}^{ano_{SS}}(t) \geqslant \theta_{SS} \tag{22}$$

Alternatively, the individual test frames are considered anomalous or normal based on the associated regularity score by using the condition mentioned in the (23). Here, θ_{SS} is the set threshold for the anomaly detection corresponding to the single scene. There is a high possibility of getting false alarms when the θ_{SS} value is very low. Conversely, there is a high possibility of missing out on the real anomalies when the θ_{SS} value is set at very high. Therefore, the θ_{SS} is set carefully for the desired application and sensitivity level.

$$S_{test_{jk}}^{reg_{SS}}(t) \leqslant \theta_{SS} \tag{23}$$

Anomaly score $S_{test_{jk}}^{ano_{SS}}$ and regularity score $S_{test_{jk}}^{reg_{SS}}$ are complimentary in nature and both are related as mentioned in (24).

$$S_{test_{jk}}^{reg_{SS}} = 1 - S_{test_{jk}}^{ano_{SS}} \tag{24}$$

Many video anomaly datasets are available that are suitable for video anomaly detection using SSF, such as Subway Entrance and Exit [41], UCSD Pedestrian [31], CUHK Avenue [42], Street Scene [1], and so on. Further, a good number of video anomaly detection works based on SSF such as [1, 13, 43–45], etc. have been reported. However, there is requirement of lightweight, efficient and robust video anomaly detection techniques for the detection and localization of the video anomalies corresponding to the SSF.

3.2 Video anomaly detection using multiple scene formulation

In the case of video anomaly detection using MSF, it is not necessary that all the normal videos must come from the same scene. Instead, normal video data coming from different scenes are used to train a single model. Subsequently, the trained model is used to detect the video anomalies in the test videos that may come from any of the trained scenes [1]. However, the video anomaly detector is able to detect the restricted varieties of anomalous events as multiple scenes are used to define the normal events. For example, video anomalies such as pedestrian walking in a restricted area, jaywalking, etc., that are treated as anomalous in a specific area of a particular scene area are excluded. This is because the same spatial region across the multiple scenes may not be the restricted zone.

Mathematically, let us consider that the complete video dataset V_{MS} is collected from the different scenarios. Hence, number of scenes available in the V_{MS} is more than one. In other words, total number of scenes being considered in the video anomaly detection using multiple scene formulation is more than one,i.e., $N_{MS} > 1$. Hence, complete video dataset V_{MS} and total number of video clips available in the V_{MS} are presented in (25) and (26), respectively.

$$V_{MS} = V^{MS}_{train} \bigcup V^{MS}_{test} \tag{25}$$

$$N^{MS}_{V} = N^{MS}_{train} + N^{MS}_{test} \tag{26}$$

Further, the training video dataset V^{MS}_{train} may be comprised of multiple video clips as represented in (3). Here, $v^{MS}_{train_j}$ represents the j^{th} video clip of the training video set V^{MS}_{train} and N^{MS}_{train} is the number of video clips available in the V^{MS}_{train}.

$$V^{MS}_{train} = \bigcup_{i=1}^{N_{MS}} \left\{ \bigcup_{j=1}^{N^{MS}_{train}} v^{MS}_{train_{ij}} \right\} = \bigcup_{i=1}^{N_{MS}} \left\{ v^{MS}_{train_{i1}}, v^{MS}_{train_{i2}}, v^{MS}_{train_{i3}}, \ldots, v^{MS}_{train_{iN^{MS}_{train}}} \right\} \tag{27}$$

Each $v^{MS}_{train_{ij}}$ comprises multiple individual frame $f^{MS}_{train_k}$ as expressed in (4). Here, M^{MS}_{train} is the total number of the frames available in each $v^{MS}_{train_j}$.

$$v^{MS}_{train_{ij}} = \bigcup_{k=1}^{M^{MS}_{train}} f^{MS}_{train_k} = \left\{ f^{MS}_{train_1}, f^{MS}_{train_2}, f^{MS}_{train_3}, \ldots, f^{MS}_{train_{M^{MS}_{train}}} \right\} \tag{28}$$

Now, V^{MS}_{train} can be rewritten by combining (27) and (28) as expressed in (29) and 30.

$$V^{MS}_{train} = \bigcup_{i=1}^{N_{MS}} \left\{ \bigcup_{j=1}^{N^{MS}_{train}} \left\{ \bigcup_{k=1}^{M^{MS}_{train}} f^{MS}_{train_{ijk}} \right\} \right\} \tag{29}$$

$$\Rightarrow V^{MS}_{train} = \bigcup_{i=1}^{N_{MS}} \left\{ \bigcup_{j=1}^{N^{MS}_{train}} \left\{ f^{MS}_{train_{ij1}}, f^{MS}_{train_{ij2}}, f^{MS}_{train_{ij3}}, \ldots, f^{MS}_{train_{ijM^{MS}_{train}}} \right\} \right\} \tag{30}$$

Similarly, the testing video dataset V^{SS}_{test} consists of multiple video clips as represented in (31). Here, $v^{MS}_{test_j}$ represents the j^{th} video clip of the testing video set V^{MS}_{test} and N^{MS}_{test} is the number of video clips available in the V^{MS}_{test}.

$$V^{MS}_{test} = \bigcup_{i=1}^{N_{MS}} \left\{ \bigcup_{j=1}^{N^{MS}_{test}} v^{MS}_{test_{ij}} \right\} = \bigcup_{i=1}^{N_{MS}} \left\{ v^{MS}_{test_{i1}}, v^{MS}_{test_{i2}}, v^{MS}_{test_{i3}}, \ldots, v^{MS}_{test_{iN^{MS}_{test}}} \right\} \tag{31}$$

Each $v_{test_{ij}}^{MS}$ comprises multiple individual frame $f_{test_k}^{MS}$ as expressed in (32). Here, M_{test}^{MS} is the total number of the frames available in each $v_{test_j}^{MS}$.

$$v_{test_{ij}}^{MS} = \bigcup_{k=1}^{M_{test}^{MS}} f_{test_k}^{MS} = \left\{ f_{test_1}^{MS}, f_{test_2}^{MS}, f_{test_3}^{MS}, \ldots, f_{test_{M_{test}^{MS}}}^{MS} \right\} \tag{32}$$

Now, V_{test}^{MS} can be rewritten by combining (31) and (32) as expressed in (33) and (34).

$$V_{test}^{MS} = \bigcup_{i=1}^{N_{MS}} \left\{ \bigcup_{j=1}^{N_{test}^{MS}} \left\{ \bigcup_{k=1}^{M_{test}^{MS}} f_{test_{ijk}}^{MS} \right\} \right\} \tag{33}$$

$$\Rightarrow V_{test}^{MS} = \bigcup_{i=1}^{N_{MS}} \left\{ \bigcup_{j=1}^{N_{test}^{MS}} \left\{ f_{test_{ij1}}^{MS}, f_{test_{ij2}}^{MS}, f_{test_{ij3}}^{MS}, \ldots, f_{test_{ijM_{test}^{MS}}}^{MS} \right\} \right\} \tag{34}$$

Similar to the SSF, the normal class distribution D_N can be estimated by using the training samples V_{train}^{MS} consisting of only normal classes from multiple scenes by modeling an automatic representation (Video Anomaly Detector) VAD_{MS} that minimizes the reconstruction cost. Mathematically, the objective function for the video anomaly detection can be formulated as represented in (35).

$$minimize \ \left\| f_{train_{ijk}}^{MS} - VAD_{MS}\left(f_{train_{ijk}}^{MS} \right) \right\|^2 \tag{35}$$

subject to constraints

$$1 \leqslant i \leqslant N_{MS} \tag{36}$$

$$1 \leqslant j \leqslant N_{train}^{MS} \tag{37}$$

$$1 \leqslant k \leqslant M_{train}^{MS} \tag{38}$$

Once the model is learned, the testing frame sequences $f_{test_{jk}}^{MS}$ are fed into the trained model VAD_{MS} to reconstruct the frames $\hat{f}_{test_{jk}}^{MS}$ as compressed in (39).

$$\hat{f}_{test_{ijk}}^{MS} = VAD_{MS}\left(f_{test_{ijk}}^{MS} \right) \tag{39}$$

The reconstruction error for a given pixel with an intensity I at a spatial location (x, y) in a particular testing frame $f_{test_{ijk}}^{MS}$ at time instant t can be calculated by using the learned model VAD_{MS} by using (41) [39].

$$e_{test_{ijk}}^{MS}(x, y, t) = \left\| I_{test_{ijk}}^{MS}(x, y, t) - \hat{I}_{test_{ijk}}^{MS}(x, y, t) \right\|^2 \tag{40}$$

$$= I_{test_{ijk}}^{MS}(x, y, t) - VAD_{MS}\left(I_{test_{ijk}}^{MS}(x, y, t) \right) \tag{41}$$

Here, $I_{test_{ijk}}^{MS}(x, y, t)$ and $\hat{I}_{test_{ijk}}^{MS}(x, y, t)$ are the pixel intensities at the spatiotemporal location (x, y, t) corresponding to the frames $f_{test_{ijk}}^{MS}$ and $\hat{f}_{test_{ijk}}^{MS}$, respectively. Further, the reconstruction error of the particular test frame at time t , $i.e.$, $e_{test_{ijk}}^{reconstruct_{MS}}(t)$ can be calculated from the known pixel-level reconstruction errors $e_{test_{ijk}}^{MS}(x, y, t)$ by using (42) [39].

$$e_{test_{ijk}}^{reconstruct_{MS}}(t) = \sum_{x,y} e_{test_{ijk}}^{MS}(x, y, t) \tag{42}$$

Subsequently, the anomaly score $S_{test_{ijk}}^{ano_{MS}}(t)$ and the regularity score $S_{test_{ijk}}^{reg_{MS}}(t)$ for the test frame $test_{ijk}^{MS}(t)$ can be calculated by using (43) [40] and (44) [39]. Here, the values of both anomaly and regularity scores are lay in the range of 0 to 1.

$$S_{test_{ijk}}^{ano_{MS}}(t)=\frac{e_{test_{ijk}}^{reconstruct_{MS}}(t)-e_{test_{ijk}}^{reconstruct_{MS_{min}}}(t)}{e_{test_{ijk}}^{reconstruct_{MS_{max}}}(t)} \tag{43}$$

$$S_{test_{ijk}}^{reg_{MS}}(t)=1-S_{test_{ijk}}^{ano_{MS}}(t) \tag{44}$$

Finally, the individual test frames are considered anomalous or normal based on the associated anomaly score by using the condition mentioned in the (45).

$$S_{test_{ijk}}^{ano_{MS}}(t)\geqslant\theta_{MS} \tag{45}$$

Alternatively, the individual test frames are considered anomalous or normal based on the associated regularity score by using the condition mentioned in the (46).

$$S_{test_{ijk}}^{reg_{MS}}(t)\leqslant\theta_{MS} \tag{46}$$

Anomaly score $S_{test_{ijk}}^{ano_{MS}}$ and regularity score $S_{test_{ijk}}^{reg_{MS}}$ are complimentary in nature and both are related as mentioned in (47).

$$S_{test_{ijk}}^{reg_{MS}}=1-S_{test_{ijk}}^{ano_{MS}} \tag{47}$$

Many video anomaly datasets are available that are suitable for video anomaly detection using MSF, such as UMN [46, 47], BEHAVE [48], Live Videos [49], UCF crime [7], ShanghaiTech Campus [50], Large-scale Anomaly Detection [51], and so on. Further, a good number of video anomaly detection works based on MSF such as [49], UCF crime [7], ShanghaiTech Campus [50], Large-scale Anomaly Detection [51], UBnormal [52], etc. have been reported. However, there is requirement of lightweight, efficient and robust video anomaly detection techniques for the detection and localization of the video anomalies corresponding to the MSF.

Notations and corresponding meanings for the VAD using MSF are presented in Table 1. As VAD using SSF is a special case of VAD using MSF where $N_{MS}=1$, these notations can be interpreted similarly where MS will be replaced by SS for the VAD using SSF. Finally, video anomaly detection using SSF and MSF are qualitatively two different problems. Depending on the targeted applications, computational infrastructure, and video anomaly datasets, video anomalies can be detected using either SSF or MSF.

4 Classification of the datasets for the video anomaly detection and localization

The availability of the bench-marked datasets is one of the significant factors that control the advancement of research for video anomaly detection and localization. Video anomaly detection and localization bench-marked datasets are scarce because of the comparatively new research field, the rarity, and the limitless variety of video anomalies (anomalous activities or anomalous events) in real-world circumstances. The limited available bench-marked datasets restrict the scope of the research problem as well as potential applications. The quantity and quality of the datasets used in the development significantly control the desired performance of the developed video anomaly detectors. Further, the use of common bench-marked video

Table 1 Notations for the multiple scene formulation to detect the video anomalies

Notation	Meaning
V_{MS}	Complete video anomaly dataset, which is collected from the different scenarios .
V_{train}^{MS}	Training video clips of multiple video clips V_{MS}.
V_{test}^{MS}	Testing video clips of multiple video clips V_{MS}.
N_{V}^{MS}	Total number of video clips available in V_{MS}.
N_{train}^{MS}	Total number of video clips available in V_{train}^{MS}.
N_{test}^{MS}	Total number of video clips available in V_{test}^{MS}.
N_{MS}	Total number of scenes being considered for VAD using MSF.
$v_{train_{j}}^{MS}$	j^{th} video clip of the training video set V_{train}^{MS}.
$v_{train_{ij}}^{MS}$	j^{th} video clip of the training video set V_{train}^{MS} for i^{th} scene.
$f_{train_{ijk}}^{MS}$	k^{th} frame of j^{th} video clip corresponding to the training video set V_{train}^{MS} for i^{th} scene.
$v_{test_{j}}^{MS}$	j^{th} video clip of the testing video set V_{test}^{MS}.
$v_{test_{ij}}^{MS}$	j^{th} video clip of the training video set V_{test}^{MS} for i^{th} scene.
$f_{test_{ijk}}^{MS}$	k^{th} frame of j^{th} video clip corresponding to the training video set V_{test}^{MS} for i^{th} scene.
D_{N}	Normal class distribution
$\hat{f}_{train_{ijk}}^{MS}$	k^{th} reconstructed frame of j^{th} video clip corresponding to the training video set V_{train}^{MS} for i^{th} scene.
$\hat{f}_{test_{ijk}}^{MS}$	k^{th} reconstructed frame of j^{th} video clip corresponding to the training video set V_{test}^{MS} for i^{th} scene.
VAD_{MS}	Video anomaly detector corresponding to MSF
$e_{test_{ijk}}^{MS}\left(x, y, t\right)$	The reconstruction error for a given pixel with an intensity I at a spatial location (x, y) in a particular testing frame $f_{test_{ijk}}^{MS}$ at time instant t.
$I_{test_{ijk}}^{MS}\left(x, y, t\right)$	Pixel intensity at the spatiotemporal location (x, y, t) corresponding to the frames $f_{train_{ijk}}^{MS}$
$\hat{I}_{test_{ijk}}^{MS}\left(x, y, t\right)$	Pixel intensities at the spatiotemporal location (x, y, t) corresponding to the reconstructed frames $\hat{f}_{train_{ijk}}^{MS}$
$e_{test_{ijk}}^{reconstruct_{MS}}\left(t\right)$	Reconstruction error of the particular test frame at time t
$S_{test_{ijk}}^{ano_{MS}}\left(t\right)$	Anomaly score for the test frame $test_{ijk}^{MS}\left(t\right)$
$S_{test_{ijk}}^{reg_{MS}}\left(t\right)$	Regularity score for the test frame $test_{ijk}^{MS}\left(t\right)$
$e_{test_{ijk}}^{reconstruct_{MS_{min}}}\left(t\right)$	Minimum $e_{test_{ijk}}^{reconstruct_{MS}}\left(t\right)$ for the particular test frame sequence.
$e_{test_{ijk}}^{reconstruct_{MS_{max}}}\left(t\right)$	Maximum $e_{test_{ijk}}^{reconstruct_{MS}}\left(t\right)$ for the particular test frame sequence.
θ_{MS}	Anomaly threshold being selected for the VAD.

anomaly datasets helps to compare the video anomaly detection algorithms fairly and selects the best one for a particular application. Though quite a few numbers of publicly available bench-marked video anomaly datasets are there, we will classify them based on their suitability for video anomaly detection using either SSF or MSF. Subsequently, the pros and cons of individual video anomaly datasets will be discussed to provide a broader perspective and in-depth knowledge.

4.1 Video anomaly datasets for single scene formulation

Video anomaly datasets for single scene formulation, i.e., single scene video anomaly datasets are comprised of videos corresponding to a single scene only [1]. In other words, all the normal and anomaly video samples are recorded at a single scene with fixed experimental setups. The detection accuracy of the models trained on the single-scene dataset is high only when inference is carried out on the same scene only. However, accuracy significantly degrades when the developed model is tested at a different scene. Hence, the generalization ability of the models trained using the single scene dataset is less as compared to that of the models trained on the multiple scene datasets. Following are the important publicly available video anomaly datasets that are suitable for video anomaly detection using SSF.

4.1.1 CAVIAR dataset

CAVIAR (Context-Aware Vision using Image-based Active Recognition) dataset [53] is primarily targeted for activity recognition. However, this dataset can be used for anomaly detection to a certain extent. It comprises 28 video sequences in total or about 26419 frames in total. The video sequences have a frame resolution of 384 x 288, frames per second (fps) of 25, and are compressed using MPEG2. All the videos are recorded in two scenarios, such as the INRIA Labs' entrance lobby at Grenoble, France (captured for CAVIAR project) and a shopping center in Lisbon. For anomaly detection purposes, walking, browsing, and meeting can be treated as normal activities, whereas collapse, leaving objects, and fighting can be treated as abnormal activities. Here, the video sequences are annotated both for target position and activities.

4.1.2 Subway dataset

The subway dataset [41] is collected from stationary surveillance cameras at the entrance and exit gates of a subway station. Here, the entrance gate video stream is 96 minutes long, and it comprises 1444249 frames in total. Similarly, the exit gate video stream is of 43 minutes long and comprises 64901 frames in total. The frame resolution of both the video streams is 512 × 384. Here, there are no separate video clips for training and testing purposes. Thus, the first 15 minutes of the video streams are usually used for training, and the rest portion is used for testing purposes. The training portion of the video streams contains only normal events, whereas the test portion of the video streams contains both normal and abnormal events. Here, both the video streams have frame-level ground truth annotations indicating whether a particular frame is anomalous or not with the help of the corresponding binary flag. In the case of the subway entrance video stream, the activity of going down the turnstiles for entering the platform is known as normal activity. This entrance video contains 66 number of abnormal activities such as walking people in the wrong direction, regular interaction among the people and fast running with sudden stopping. Similarly, in the case of the subway exit video stream, the activity of exiting from the platform, passing through the turnstiles, and turning to the right at the top of the stairs is known as normal activity. This exit video contains 19 number of abnormal activities such as walking for exit in the wrong direction and loitering near the exit gate of the subway. The dataset provides one crucial challenge of the detection of video anomalies in practical scenarios in real-time. Here, no spatial ground truth is presented, and hence, performance evaluation at the spatial level is not possible.

4.1.3 MIT Traffic dataset

The MIT Traffic dataset [28, 54, 55] consists of challenging video sequences from the crowded road traffic scenes. The dataset has 90 minutes length, a spatial resolution of 720 × 480, and fps of 30 correspondings to a street corner. Here, though the traffic flow is relatively less busy than other street intersection datasets, it is less regulated and more complicated due to the presence of two types of anomalies in a single clip [8].

4.1.4 PETS dataset

The PETS 2009 dataset comprises both normal events and abnormal events corresponding to the crowd walking and crowd escaping, respectively [56, 57]. The frame sequences corresponding to the individuals walking in different directions are used to extract the training and normal testing samples. Further, the frame sequences corresponding to people walking and running in one direction are used to extract the anomalous frames meant for testing. Anomalous events such as walking in the wrong direction, running, multiple flows of the crowd, sudden dispersion, and splitting of the crowd are treated as video anomalies. The resolution of this dataset is 576 × 84. Here, maintaining the video anomaly detector's effectiveness across scenes having variable crowd density is one of the crucial research challenges.

4.1.5 U-turn dataset

The U-turn dataset [58] is collected by a stationary surveillance camera focused on a traffic junction. The video clip consists of 6057 frames in total. Here, regular traffic such as trams passing, cars driving in different directions, and pedestrians walking is considered normal events. However, abnormal traffic incidents such as illegal U-turns by cars, the person dropping a bag, and abandoning it are known as video anomalies.

4.1.6 QMUL junction dataset

QUML Junction dataset [59–62] is created by capturing videos at an fps of 25 using a stationary surveillance camera from the busy street intersections. Here, the traffic light is used to control the three traffic flows in different directions. The total length of the dataset is 60 minutes comprising 89999 frames having a spatial resolution of 360 × 288. The dataset contains various anomalous events such as illegal U-turns, traffic interruption by emergency vehicles, and so on. No official ground truth, training, and testing partitions are provided. Hence, ground truth may be extracted by using manual labeling from the test video sequences. Similarly, splitting the dataset into training and testing sets can be carried out based on the desired events. The dataset provides various research challenges such as complex interactions among vehicles and pedestrians, changing complexity due to changes in traffic flow, illumination changes, shadow effects, and video capturing noise [8].

4.1.7 UCSD pedestrian dataset

A static or stationary surveillance camera is used to acquire the UCSD Pedestrian dataset [31] at an fps of 10 from an elevation covering an outdoor scene, namely the pedestrian walkways. The UCSD pedestrian dataset comprises two subsets of the dataset known as Ped1 and Ped2, corresponding to two different walkways. Ped1 has 70 total video clips, i.e., the number of training and testing video clips is 34 and 36, respectively. Further, each video clip of the Ped1 contains 200 frames. Similarly, Ped2 has 28 total video clips, i.e., the number of training and testing video clips is 16 and 12, respectively. The training clips consist of only normal events, i.e., only pedestrians. However, the testing clips consist of both normal events and video anomalies [63]. Ped1 and Ped2 contain 14000 frames with 40 video anomalies and 4560 frames with 12 video anomalies, respectively. The significant difference between these two subsets is that Ped1 has perspective distortions due to walking of pedestrians towards and away from the capturing camera, whereas Ped2 contains no perspective distortion [64]. The video anomalies are caused due to the circulation of the non-pedestrian objects and abnormal motion patterns of the pedestrians in the walkways. Hence, all other objects such as cars, bikers, skaters, and vehicles are treated as video anomalies apart from pedestrians. All the testing video clips have frame-level ground-truth annotations, whereas only ten testing clips have pixel-level ground truth annotations. Usually, most of the video anomaly detection methods perform relatively better on Ped2 as compared to Ped1. The potential reason may be that the complexity and variance of the crowd density for Ped2 are less than that of Ped1 [65]. Illumination variation, variable crowd density ranging from sparse to very crowded, scale changing of the objects, and perspective distortion are the essential research challenges provided by the UCSD Pedestrian dataset. However, this dataset is of modest size in terms of the number of frames and varieties of anomalies.

4.1.8 Pedestrian crossing dataset

Pedestrian Crossing Dataset [66] consists of a video of 45 minutes that caputures a pedestrian activity in a street intersection. The pedestrian's behavior while crossing the traffic flow as per the traffic rule is considered a normal event. Any illegal movement of the pedestrian is known as an abnormal event. The video is captured at an fps of 25 with frame resolution of 360 × 288.

4.1.9 CUHK avenue dataset

The CUHK Avenue dataset [42] is collected from a fixed surveillance camera in real scenarios of CUHK campus avenue. It contains 16 training video clips consists of 30652 frames and 21 testing video clips consists of 15324 frames. The total number of frames present in the dataset is 30652, and the individual frame resolution is 640 × 360. The total duration of the dataset is 30 minutes. The training video clips have only normal events, whereas the testing video clips consist of both normal and abnormal events. There are forty-seven number of abnormal events present in the testing video clips. The abnormal events or video anomalies are mostly caused due to two varieties of activities, such as the movement of non-pedestrian entities in the walkways and abnormal motion patterns executed by the pedestrians. Bikers, skaters, small carts, walking across the walkways or in the surrounding grass are the frequently occurring video anomalies. Further, occasionally video anomalies such as wheelchairs are also recorded. Here, all the anomalies occur naturally and hence, provide a realistic environment. In other words, the anomalies are not staged or synthesized for the creation of datasets. Object-level ground truth annotations are provided for all the testing clips, i.e., video anomalies are labeled in a spatial location with rectangles. Further, these annotations can be used to evaluate the performance of the video anomalies at frame-level, and pixel-level as frames having marked rectangles can be used to get frame-level annotations, and spatial location of the rectangle can be used to obtain approximate pixel-level annotations. A slight camera shaking is introduced in the test video clips to provide some research challenges.

4.1.10 A day on campus dataset

A Day on Campus (ADOC) dataset [67] is suitable for video anomaly detection using single scene formulation. This dataset overcomes the important disadvantages of the existing datasets by including more number of anomalous events in the crowded scenarios with varying illumination conditions, background clutters, and occlusions. The dataset is challenging as there are no separate categories of classes for anomalous and normal events. Here, any event that occurs very few times (less probable events) is considered a video anomaly. Further, a total 721 number of anomalous events are annotated using both manual and automatic approaches. Subsequently, ground truths are available in the form of Bounding Boxes (BB). There are 25 categories of events included in this ADOC dataset. These classes are riding a bike, walking on grass, driving a golf cart, walking with the suitcase, having a conversation, riding a skateboard, birds flying, walking with a bike, pushing a cart, person vending, standing on a walkway, bending, walking a dog, riding a mobility scooter, running, group of people, cat or dog, crowd gathering, holding a sign, walking with balloons, a bag left behind, truck on a walkway, a person on a knee scooter, camera overexposure, and person smoking. The dataset is challenging due to the presence of large varieties of classes, scenes having varying illuminations, background clutters, and occlusions.

4.1.11 Street scene dataset

Street Scene dataset [1] is suitable for video anomaly detection using single scene formulation. This dataset is built to overcome the important disadvantages of the existing datasets. These disadvantages include simplicity of scenes, a small number of anomalies, lack of diversity in anomalies, very low resolutions of some datasets, presence of scripted as well as staged anomalies in some datasets, inconsistency in annotations, and lack of pixel-level as well as frame-level ground truth annotations in most of the datasets. The dataset is captured by a stationary USB camera that looks down on a scene that comprises a two-lane street for bike lanes and pedestrian sidewalks. Further, the dataset consists of 46 training video clips and 35 testing video clips. These videos are collected during the daytime only at various times during two consecutive summers. It is comprising of a total of 203257 frames having individual spatial resolution 1280 × 720 at an fps of 15. Total frames are divided into two subsets, such as 56847 meant for training and 146410 meant for testing. Pixel-level ground truth annotations are provided in the form of bounding boxes around each anomalous event in each frame of the testing video. Here, each annotation box is labeled with a tracking number. Hence, a single frame may have more than one bounding box. There are 17 classes of the video anomalies such as jaywalking, biker outside lane, loitering, dog on the sidewalk, a car outside lane, workers in buses, biker on the sidewalk, pedestrian reveres direction, car U-turn, car illegally parked, person opening trunk, a person exists car on the street, a skateboarder in the bike lane, person sitting on the bench, meter maid ticketing car, a car turning from the parking place, and motorcycle drives onto the sidewalk. The dataset is challenging due to the presence of large varieties of anomalous classes, scenes having changing shadows, and moving backgrounds such as blowing a flag as well as trees in the wind.

4.2 Video anomaly datasets for multiple scene formulation

Video anomaly datasets for multiple scene formulation, i.e., multiple scene video anomaly datasets are comprised of videos corresponding to multiple scenes [1]. In other words, all the normal and anomaly video samples are recorded at different scenes with fixed or varying experimental setups. The detection accuracy of the models trained on the multiple-scene dataset is slightly low compared to those trained on single-scene datasets only when inference is carried out on a particular scene only. However, accuracy does not degrade significantly when the developed model is tested at different scenes. Hence, the generalization ability of the models trained using the multiple-scene dataset is high as compared to that of the models trained on the single-scene datasets. Following are the important publicly available video anomaly datasets that are suitable for video anomaly detection using MSF.

4.2.1 UMN dataset

UMN dataset comprises both normal and abnormal events such as various panic-driven or crowd escape events recorded on the University of Minnesota campus [46, 47]. The dataset consists of three different scenarios of both indoor and outdoor scenes, such as lawn, indoor, and plaza. Here, in each scenario, the normal walking of the group of people in various directions is treated as normal events, whereas the sudden run away (escape) by the same group of people is treated as abnormal events (video anomalies). All three video clips have a frame resolution of 320 × 240. The video clips corresponding to lawn, indoor, and plaza have 1450, 4415, and 2145 frames, respectively. Here, there are no separate video clips or frame sequences are present for both training and testing. Instead, the anomalous frames are extracted from the common frame sequence for testing purposes. Handling of occlusion is one of the important research challenges that is provided by this dataset.

4.2.2 i-Lids dataset

This dataset was developed for performance evaluation of detection and tracking algorithms, particularly for i-Lids bag and vehicle detection challenge of IEEE conference on Advanced Video and Signal based Surveillance, 2007 [68, 69]. However, it can be used to detect anomalous objects such as abandoned baggage in the platform scenario and wrongly parked vehicles in the road traffic. The abandoned baggage and wrongly parked vehicle correspond to the indoor and outdoor scenes, respectively. Intentionally introduced occlusion is one of the significant challenges provided by this dataset. The dataset comprises seven videos that are recorded at an fps of 25 with a frame resolution of 720 × 576.

4.2.3 UCF crowd segmentation dataset

The UCF crowd segmentation dataset [70–72] is developed for crowd flow and stability analysis densely crowded scenarios such as large gatherings of people at events such as religious festivals, parades, concerts, football matches, and so on. Few examples of the dataset are a scene from New York City marathon, a large crowd participating in a political rally in

Los Angeles, pilgrims circling around Kabba in Mecca, etc. Here, detection and management of abnormal or anomalous behavior in time is very much necessary for managing these large gatherings to maintain public safety. This dataset contains around 36 real-world surveillance videos covering normal activities in various situations such as platforms, shopping malls, marketplaces, religious places, etc. This can be used for training a video anomaly detector to detect anomalous crowd behaviors.

4.2.4 Web dataset

Web dataset [72] is a video anomaly dataset focused on crowd activities in urban scenarios. It comprises high-quality and documentary videos downloaded from the websites such as Getty Images and ThoughtEquity.com. The dataset consists of twelve sequences of normal crowd scenes and eight scenes of abnormal crowd scenes. The normal sequences include pedestrian walking and marathon running. Further, abnormal sequences include escape panics, protesters clashing, and crowd fighting.

4.2.5 BEHAVE dataset

The BEHAVE dataset [48] is available in either four video clips or 76800 individual frames extracted at 25 fps with a spatial resolution of 640 × 480. Here, the anomalies are in accordance with crime-oriented abnormal behaviors such as chase, fight, and run together. There are no separate files for training and testing. Hence, training and testing subsets can be segregated as per the desired event. However, anomalous activities are scripted and acted by the actors.

4.2.6 Traffic dataset

There are few important traffic datasets [73–77] for video anomaly detection. Two important traffic datasets [73, 75] comprise two video clips (each video clip is of 3-hour duration) corresponding to a crowded scene in Zurich traffic and two video clips (each clip is of 1-hour duration) corresponding to London traffic. All these video clips are recorded at an fps of 25. Similarly, a 50 minutes duration video with a spatial resolution 360 × 288 at fps of 30 is collected [74]. Here, the footage includes the movement of cars and people in accordance with the traffic control signal. Further, another traffic dataset, a 5-hour long surveillance video, is captured to detect video anomalies based on trajectory analysis [76]. Here, whenever people, bicycle, and vehicles follow their normal or routine path, then it is known as normal events. However, whenever the same objects follow an abnormal path, then it is known as abnormalities. This dataset comprises 1000 trajectories in total where 898 normal and 102 abnormal trajectories. However, these datasets are not suitable for deep learning approaches meant for video anomaly detection as the quantity and quality of the videos are not sufficient.

4.2.7 Anomalous behavior dataset

The Anomalous behavior dataset or York university dataset [78–81] contains eight frame sequences corresponding to egiht different scenarios such as Traffic-Train, Belleview, Boat-Sea, Boat-River, Subway-Exit, Camouflage, Airport-Wrong Direction having a total number of frames such as 19218, 2918, 450, 250, 32426, 1050, 1629, and 2200 respectively. The dataset provides a wide range of challenges such as illumination effects, scene clutter, variable target appearance, rapid motion, and camera jitter for research. Manually constructed ground truths for identifying anomalous behaviors relative to a training portion of the video are available for all the frame sequences.

4.2.8 VIRAT dataset

The VIRAT dataset [34] is primarily developed for activity recognition. However, it can also be used for context-specific anomaly detection. Variations in the activities and the presence of clutters in the scene are the two important, challenging characteristics of the dataset. The dataset consists of short video clips of variable duration from 2 to 15 minutes corresponding to different realistic scenarios. It contains around 30 number of events.

4.2.9 Violent flows dataset

Violent Flows Dataset [82] is used to detect the breaking of violence in densely crowded environments. The dataset is comprised of 246 number of video clips whose length varies between 1.04 Sec to 6.52 Sec. Mostly, the videos are collected from the real-time footage of various crow violence at football and hockey stadiums. However, the dataset is primarily meant for the classification of crowd behaviors as normal and abnormal. Only video level or folder level annotations are given. Hence, it can not be used for crowd anomaly detection with the creation of proper annotations, at least at the frame level.

4.2.10 BOSS dataset

Boss dataset [83] is collected from the nine stationary surveillance cameras mounted in a train at 25 frames per second. The videos have spatial resolution of 720 × 567 and each video last for 1 minute to 5 minutes. It contains three normal videos and eleven abnormal behavior videos containing various anomalies such as the person with a disease, grabbing cell phone, fighting, grabbing the newspaper, harassing, fainting, and panicking, which are performed by the actors as per predefined scripts. Here, the variations in the anomalies are limited, the dataset is small in size, and the number of video clips is also less.

4.2.11 LV dataset

Live Videos (LV) dataset [49] is one of the first video anomaly datasets recorded from the realistic surveillance scenarios in the true sense. The dataset contains 28 realistic video clips where the events occur naturally in a diverse subject interaction mode without the staged one as per the predefined scripts. Here, the video clips are recorded with different frame resolutions (ranging from 176 × 144 to 1280 × 720) at different fps (ranging from 7.5 to 30). The dataset contains highly unpredictable and naturally occurring video anomalies of different duration. Few anomalies are very difficult to be detected as they are of very short duration consists of the only couple of frames. The total duration of the dataset is 3.93 hours comprises 309940 frames in total (out of which 68989 frames are anomalous frames). Officially, there are no separate clips for training and testing sets. Instead, the dataset is scenario correspondence where the training and test data are captured from the same scene. The dataset is captured from various diverse and challenging scenarios such as outdoor, indoor, streets, highways, traffic intersections, and public places. These scenarios provide important research challenges of variable crowd density ranging from no subject to very crowded environments. Further, the video clips are acquired in various challenging environmental conditions, such as changing illuminations and camera motions. The dataset contains 14 varieties of video anomalies such as robberies, wrong U-turns, crowd panic, loitering, fighting, homicide, trespassing, kidnapping, fire, car driving in the wrong direction, car accidents, falling of people, vandalism, clashinf of people, thefts, loitering, hit and runs. However, the size of the dataset is 863 MB which is comparatively less in size for data-driven approaches such as deep learning-based video anomaly detection.

4.2.12 UCF crime dataset

A new large-scale dataset named UCF crime [7] is introduced for video anomaly detection and activity recognition. The UCF crime dataset consists of 1900 long and untrimmed real-world surveillance videos covering both normal activities and realistic anomalies. It contains 13 realistic anomalies such as abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism. The dataset can be used for video anomaly detection where all the anomalies and all the normal activities are arranged in two different groups. Further, the dataset can be used for the recognition of anomalous activities where all the 13 anomalous classes are split into 13 subgroups. The dataset is 96.34 GB in size and has a total of 128 hours of content. It is by far the largest dataset among the publicly available datasets for video anomaly detection and hence, more suitable for video anomaly detection using deep learning methods. The anomalies present in the videos are quite complex and natural that are collected from the real-world surveillance videos available over the Internet. Here, both anomalous and normal video clips are available for training. The videos are captured by hundreds of stationary surveillance cameras from many scenarios, and hence, the dataset provides a large set of diversity in the scenarios. Here, only weakly-labeled annotations(only video-level annotations or folder-level annotations) are available for training of the video anomaly detector. However, for evaluating the performance of the video anomaly detector during testing, temporal annotations, i.e., the start and end frames of the anomalous video segment, are provided. Hence, only frame-level evaluation is possible for this dataset. Further, the abnormal frames of this dataset are not completely labeled. Here, spatial evaluation is not possible due to absence of spatial-level annotations such as pixel-label and bounding boxes. However, this dataset seems to be more akin to activity recognition [1].

4.2.13 ShanghaiTech campus dataset

Previously, most of the datasets were lacking diversity and viewing angles as they were recorded with a single stationary camera at a fixed viewing angle corresponding to a particular scene [84]. Hence, ShanghaiTech Campus [50] has been built to increase the scene diversity and volume of data. It is one of the very challenging video anomaly datasets that contains 330 training video clips and 107 testing video clips. It consists of 130 abnormal events (video anomalies) covering 13 realistic scenes of complex lighting conditions. The total duration of the ShanghaiTech Campus dataset is approximately (App.) 208 min. Few new varieties of abnormalities such as chasing and brawling caused by sudden motion are introduced in this dataset. The dataset comprises 317398 frames in total, out of which 274515 and 42883 are used for training and testing purposes, respectively. Further, there are 300308 normal frames and 17090 anomalous frames. Also, pixel-level-ground truth annotations are provided, which helps in the performance evaluation of both anomaly detection and localization. This dataset is intended for developing a single model by following the SSF. However, the dataset is collected from multiple scenes, and hence it follows the MSF.

4.2.14 LAD dataset

Large-scale Anomaly Detection (LAD) dataset [51] is proposed to address two critical problems, i.e., limited in scale and lack of annotations for the precise anomalous time duration, in the previously existing datasets. It contains two thousand video sequences. It comprises fourteen anomaly classes: crash, crowd, destroy, drop, falling, fighting, fire, fall into the water, hurt, loitering, panic, thieving, trampled, and violence. All the videos are recorded at an fps of 25 from 1895 different visual scenarios. This dataset contains proper annotations at video levels and frame levels. Moreover, the dataset comprises of good or HD quality video sequences as video sequences having a poor resolution, incomplete, and ambiguous anomalies are dropped at the collection stage. Supervised learning-based video anomaly detection may be applied to this dataset due to the availability of the proper annotations and their large size.

4.2.15 UBnormal dataset

UBnormal dataset [52] is proposed to address two crucial issues such as lack of anomalous samples compared to normal samples and inefficient detection of new unseen video anomalies associated with the existing datasets. Further, the UBnormal dataset is a supervised open-set video anomaly dataset comprised of multiple virtual scenes where the training, validation, and testing samples encompass different anomaly classes. It comprises twenty-two anomalous events such as running, falling, fighting, sleeping, crawling, having a seizure, laying down, dancing, stealing, rotating 360°, shuffling, walking injured, walking drunk, stumbling walk, car crash, running injured, fire, smoke, jaywalking, driving outside lane, jumping, people and car accident. This dataset contains proper annotations at pixel, frame, and object levels. The diversity of the UBnormal dataset is increased by including foggy scenes, night scenes, fire scenes, smoky scenes, higher variations in anomaly types and scenes, and multiple object categories such as people, cars, skateboards, bicycles, and motorcycles. Moreover, the dataset comprises HD-quality video sequences generated at 30 fps with a minimum height of 720 pixels. Here, virtual scenes are created using 2D background images and 3D animated objects with the help of Cinema4D software. However, there may be performance degradation of the VAD systems developed using the UBnormal dataset when deployed in the field as the dataset lacks real-world anomalous samples.

5 Comparative analysis

Different datasets have been created for developing and testing video anomaly detection methods. The fewness of the benchmarked datasets available for the video anomaly detection and location is due to rareness as well as infinite varieties of the anomalous activities in real-life scenarios [15]. Comparative analysis of the benchmarked datasets for video anomaly detection based on qualitative parameters for single scene and multiple scene formulations are presented in Tables 2 and 3, respectively. In Tables 2 and 5, the comparison is carried based on the various qualitative parameters such as surveillance environment, scenarios covered, challenges offered by the datasets, anomalous events involved, and availability of the

Table 2 Comparative analysis based on qualitative parameters of the bench-marked single scene video anomaly datasets

Datasets	Environment	Scenario	Challenges	Anomalous Events	GT
CAVIAR [53]	Sparsely crowded	Indoor premises such as entrance lobby and shopping center	Appearance detection, occlusions	Interaction anomalies involving fighting, people walking together and splitting	Yes (OL)
Subway Entrance and Exit [41]	Moderately crowded	Underground train station (indoor environment)	Real-time	Avoiding turnstiles, wrong direction	Yes (FL)
MIT Traffic [28, 55]	Moderately crowded	Road traffic	Video captured by stationary camera	Detection of pedestrian as anomaly on the public road	Yes (Not useful for VAD)
U-Turn [58]	Moderately crowded	Traffic junction	Low resolution and complex background	Illegal U-turns by car, abandoning of bag by a person	No
PETS' 09 [57]	Moderately crowded	Multi-sensor sequence of various crowd activities possessing calibration data	Variable crowd density	Walking, running, multiple flows of crowd, sudden dispersion, splitting	Yes (FL)
QMUL Junction [59–62]	Moderately crowded	Covers the traffic intersection at the junction	Poor quality	Wrong direction of vehicles	No
UCSD Ped 1 and Ped 2 [31]	Sparsely and moderately crowded	Pedestrian walking on walkway and street	Illumination variations	Skaters, carts and vehicles on pedestrian walkway	Yes (FL, PL)

Table 2 continued

Datasets	Environment	Scenario	Challenges	Anomalous Events	GT
Pedestrian Crossing [66]	Moderately crowded	Street intersection	Complex road scene	Traffic rule violation	No
CUHK Avenue [42]	Sparsely to moderately crowded	Covers traffic and pedestrian scenarios	Variation in crowd density	Unusual behavior, walking, in wrong direction, unattended object	Yes(FL, PL)
Street Scene [1]	Moderately and densely crowded	A two-lane street having bike lanes and pedestrian sidewalks	Occurrence of multiple activities such as cars driving, turning, stopping, and parking; pedestrians walking, jogging, and pushing strollers; and bikers riding in bike lanes. Changing shadows and presence of moving objects such as flag and trees blowing due to wind in the backgrounds	Pedestrians performing jaywalking,loitering, walking in opposite direction; bikers on sidewalks and outside of the bike lanes; cars making u-turns, cars parked illegally, cars outside a car lane.	Yes (FL, PL)
ADOC [67]	Sparsely, moderately, and densely crowded	Large university campus	Crowded scenes with varying illuminations, background clutters, and occlusions	Low-frequency events such as person on knee scooter, person smoking, camera overexposure, etc.	Yes (FL, PL, BB)

Table 3 Comparative analysis based on qualitative parameters of the bench-marked multiple scene video anomaly datasets

Datasets	Environment	Scenario	Challenges	Anomalous Events	GT
UMN [47, 87]	Sparsely and moderately crowded	Indoor and outdoor premises	Occlusion	Abandoned objects, unusual crowd behavior, camera sabotage, intrusion, loitering	Yes (FL)
i-Lids [68, 69]	Sparsely crowded	Abandoned object and parking	Small size	Abandoned objects (baggage), parking of vehicle in forbidden area	Yes (FL)
UCF Crowd Segmentation [70–72]	Densely crowded	Religous festivals, parade, concerts, football matches, political rallies	small size objects, crowd anomaly detection	abnormal crowd behavior	No
Web [72]	Moderately and densely crowded	Urban scenarios	Low resolution	Escape panics, protesters clashing, and crowd fighting	No
Behave [48]	Sparsely crowded	Street, parking area	Crime oriented anomaly detection	Chase, fight, and run-together	Yes (FL)
Traffic [73–77]	Moderately crowded	Road	Crowd density variation, complex background	When object follows an abnormal path	Yes (Not useful for VAD)
Anomalous Behavior / York [79]	Moderately crowded	Train station	Variation in illumination and motion, clutter, camera jitter, variable target appearance	Boarding on and off the train, wrong direction	Yes (FL)
Violent Flows [82]	Densely Crowded	Both violent and non-violent activities of crowd at public places like stadium	Collected from YouTube	Crowd violence	No
BOSS [83]	Sparsely to densely crowded	Inside a metro train	Motion and stabilization in video	Harass, disease, panic	Yes (VL)

Table 3 continued

Datasets	Environment	Scenario	Challenges	Anomalous Events	GT
Live Videos (LV) [49]	Sparsely and moderately crowded	indoor/ outdoor premises, traffic intersections, roadways, public areas	Real-world surveillance videos	Robberies, wrong U-turns, crowd panic, loitering, fighting, homicide, trespassing, kidnapping, fire, car driving in wrong direction, car accidents, falling of people, vandalism, people clashing, thefts, loitering, hit and runs	Yes (FL, PL)
UCF-Crime [7]	Sparsely, moderately, and densely crowded	Real-world anomalies	Realistic anomalies	Abuse, arrest, arson, assault, accident, burglary, explosion, fighting, robbery, shooting, stealing, shoplifting, and vandalism	Yes (VL, FL)
ShanghaiTech Campus [50]	Sparsely to moderate crowded	Covers diverse anomalous scenes (approx. 13)	Captured from multiple cameras with different view angles under varying illumination conditions	Suspicious activities characterized by violent motions like brawling, chasing, skaters, bikers and trolley on the pedestrian walkways	Yes (FL, PL)
LAD [51]	Sparsely, moderately, and densely crowded	Diverse scenarios including indoor and outdoor scenes such as road, stadium, railway station, shopping mall, play ground, office, etc.,	Varieties of real-world scenarios with varying crowd density, varying background complexity, changing environmental conditions	Crash, crowd, destroy, drop, falling, fighting, fire, fall into water, hurt, loitering, panic, thieving, trampled, and violence	Yes (VL, FL)
UBnormal [52]	Sparsely and moderately crowded	Diverse scenarios including indoor and outdoor scenes such as road, stadium, railway station, airport, market street, office rooms, etc.,	Varieties of virtual scenes that are very much similar real-world scenarios with varying crowd density, involvement of multiple objects participating in the anomalous events, varying background complexity, changing environmental conditions like foggy and night scenes	Running, falling, fighting, sleeping, crawling, having a seizure, laying down, dancing, stealing, rotating 360°, shuffling, walking injured, walking drunk, stumbling walk, people and car accident, car crash, running injured, fire, smoke, jay walking, driving outside lane, and jumping	Yes (PL, FL, OL)

ground truth (GT). Moreover, GTs are provided at various levels such as Object Level (OL), Video Level (VL), Frame Level (FL), pixel Level (PL), and Bounding Boxes (BB). Similarly, Comparative analysis of the benchmarked datasets for video anomaly detection based on quantitative parameters for single scene and multiple scene formulations are presented in Tables 4 and 5, respectively. Further, in Tables 4 and 5, the comparison is carried out based on the various quantitative parameters such as dataset duration (Dur.), size, resolution (Reso.), recording speed in fps, number of scenes covered (S_n), types/classes of anomalous events (A_{cls}), number of anomalous instances (A_{inst}), the total number of video clips available (V_{tot}), the number of training video clips (V_{trn}), the number of testing video clips (V_{tst}), the total number of frames available (F_{tot}), number of frames available for training (F_{trn}), number of frames available for testing (F_{tst}), number of regular or normal frames available (F_{reg}), and number of Irregular or abnormal frames available (F_{irreg}). The selection of a good combination of test data is one of the crucial components in deep learning-based developments [85]. The majority of the publicly available datasets contain simulated abnormal behaviors, a limited number of realistic anomalous behaviors, videos that are recorded using predefined scripts, training and test samples from different camera setups, videos mostly on ideal environment [49]. Deep learning-based video anomaly detection methods require large datasets covering realistic anomalous behaviors. Recently, few important datasets such as the LV dataset [49], ShanghaiTech Campus dataset [50], UCF-Crime datasets [7], Street Scene [1], and LAD [51] are put forth for developing as well as testing deep learning-based video anomaly detection methods. However, there are nine most widely video anomaly detection datasets such as UMN [46, 47], Subway [41], UCSD Pedestrian [31], CUHK Avenue [42], [49], ShanghaiTech Campus dataset [50], UCF-Crime datasets [7], Street Scene [1], and LAD [51]. This is due to their easy availability, presence of diversity in scenarios, good quality, and quantity.

Further, based on the comparative analysis, it is evident that the methodology for selecting video anomaly datasets to achieve higher performances is a complex process. The methodology for selecting appropriate video anomaly datasets depends on the context, targeted application, desired accuracy, available time, and computational resources. However, a generic yet effective guideline can be framed to select the best video anomaly datasets for the problem in hand as follows.

- Defining the research objective: The research objective must be well-defined in terms of desired accuracy and latency for the target application. Further, it must include a comprehensive list of potential anomalies with their context dependency for better planning to curb the number of false alarms and miss detection.
- Data availability: Based on the research objective, the availability of the video anomaly datasets must be checked. Further, suitable videos may be collected from the targeted surveillance zones and subsequently preprocessed as well as annotated to develop a large corpus of suitable video anomaly datasets that can be used for building efficient VAD models.
- Data volume: As developing efficient and robust AI-based models for the detection and localization of the video anomalies is a data-driven approach, the volume of the video anomaly datasets must be high.
- Data quality: Along with data volume, data quality must be ensured to achieve the desired accuracy from the trained video anomaly detector. Hence, high data quality in terms of higher resolution, lower noise or artifact levels, and training samples free from corrupted frames or videos must be maintained to obtain higher accuracy.

Table 4 Comparative analysis based on quantitative parameters of the bench-marked single scene video anomaly datasets

Datasets	Dur.	Size	Reso.	fps	S_n	A_{cls}	A_{inst}	V_{tot}	V_{trn}	V_{tst}	F_{tot}	F_{trn}	F_{tst}	F_{reg}	F_{irreg}
Subway Entrance [41]	96 min	3.23 MB	512 × 384	25	1	5	66	1	-	-	136525	20000	116524	134124	2400
Subway Exit [41]	43 min	3.23 MB	512 × 384	25	1	3	19	1	-	-	72401	7500	64901	71681	720
QMUL Junction [59–62]	50 min	324 MB	360 × 288	25	1	2	-	1	-	-	89999	-	-	-	-
UCSD Ped1 [31]	23.33 min	706 MB	158 × 283	10	1	5	54	70	34	36	14000	6800	7200	9995	4005
UCSD Ped2 [31]	7.6 min	520 MB	240 × 360	10	1	5	23	28	16	12	4560	2550	2010	2924	1636
CUHK Avenue [42]	30 min	778 MB	640 × 360	-	1	5	47	37	16	21	30652	15324	15324	26832	3820
Street Scene [1]	22.5 min	45.7 GB	1280 × 720	15	1	17	205	81	46	35	20257	56847	146410	-	-
ADOC [67]	24 hours	- GB	1920 × 1280	3	1	25	721	-	-	-	259123	-	-	142962	97030

Table 5 Comparative analysis based on quantitative parameters of the bench-marked multiple scene video anomaly datasets

Datasets	Dur.	Size	Reso.	fps	S_n	A_{cls}	A_{inst}	V_{tot}	V_{trn}	V_{tst}	F_{tot}	F_{trn}	F_{tst}	F_{reg}	F_{irreg}
UMN [47, 87]	5 min	63.7 MB	320 × 240	15	3	1	11	-	-	-	3855	N/A	N/A	-	-
UCF Crowd Segmentation [70, 72]	10.48 min	142 MB	480 × 360	-	7	9	-	39	36	3	-	-	-	-	-
BEHAVE [48]	148.58 min	300 MB	640 × 480	25	2	3	-	4	-	-	76800	-	-	-	-
Anomalous Behavior [79]	20 min	149 MB	320 × 240	-	8	8	-	-	-	-	60141	-	-	-	-
LV [49]	3.93 hours	863 MB	177 × 144, 1280 × 720	7.5 -30 min	6	14	-	30	-	-	309940	127500	182440	240951	68989
UCF Crime [7]	128 hours	96.34 GB	1280 × 720	-	> 500	13	-	1900	950	950	-	-	-	-	-
ShanghaiTech Campus [50]	App. 208 min	6.68 GB	846 × 480	-	13	-	130	437	330	107	317398	274515	42883	300308	17090
LAD [51]	App. 25.12 hours	40.5 GB	320 × 240, 1280 × 720, 1920 × 1080	25	>30	14	-	2000	1000	1000	317525	-	-	-	-
UBnormal [52]	App. 132 min	15 GB	1280 × 720 (minimum)	30	29	22	660	543	268	211	236902	116087	92640	147887	89015

- Data annotation consistency: The annotation of the video anomaly datasets must be consistent throughout the of the video samples at video, frame, and pixel levels.
- Data diversity: The video anomaly datasets should be free from redundant and duplicate samples. Alternatively, video anomaly datasets must have different varieties of normal and abnormal samples to increase the robustness of the model.
- Availability of data ground truths: Any VAD model must be evaluated to measure its efficacy. Hence, the datasets must have one or multiple levels of annotations files at FL, PL, DPL, and OL to evalute the performance of the developed VAD model.
- Benchmarking: It is advisable to compare the results of the proposed model on the selected video anomaly datasets with other SOTA models for the same datasets for benchmarking purposes. Hence, video anomaly datasets that are widely used and publicly available must be selected for the research purpose.
- Data documentation: While selecting video anomaly datasets, it is essential to understand the data properly, for which proper documentation must be available. Hence, video anomaly datasets having proper self-explanatory document is widely used and cited.
- Data privacy and ethical issues: While selecting or collecting video anomaly datasets, the researchers must ensure that appropriate data privacy and ethical issues as per the rule of land have been considered.

6 Issues with existing datasets

The various important issues or problems associated with the different publicly available datasets of video anomaly detection may be outlined as follows.

6.1 Fewness of the datasets

There are only a few publicly available benchmarked datasets for video anomaly detection as this research area is picking up recently. Further, this fewness of the benchmarked datasets has resulted due to the equivocal and rareness nature of the video anomalies in practical scenarios [15, 22, 86]. Hence, there is a requirement for a quite large number of video anomaly datasets corresponding to a diverse range of applications.

6.2 Data imbalance

Video anomalies are irregular (rare) as well as abnormal patterns that can be detected and localized in spatiotemporal dimensions. Hence, getting an equal amount of positive (anomalous/ abnormal) and negative (normal) training samples corresponding to a particular anomalous event is not always possible [25]. Subsequently, a data imbalance problem exists between positive and negative data samples inherently due to a very high difference between the number of samples corresponding to both the classes. Therefore, most of the widely used datasets [1, 7, 31, 46, 47, 49, 50, 87] are also associated with the inherent data imbalance problem. This data imbalance problem precludes the supervised learning-based models for video anomaly detection using these datasets. However, One-Class Classification (OCC) involves learning the model from the normal data only and predicting the unseen data as normal or anomaly [88]. Hence, the OCC also helps in addressing the data imbalance problem.

6.3 Annotation inconsistency

Consistency in annotations is very much essential for developing a deep learning-based efficient video anomaly detector. However, most of the datasets such as Subway [41], [46, 47], UCF Crime [7], etc. suffer from annotation inconsistency. This annotation inconsistency includes spelling error/mismatch in the names of the classes, wrongly tagging of normal events as anomalous ones or vice versa, missing annotations, etc., at few places. Therefore, the performance of the developed model decreases, and also difficulty arises during the performance evaluation of the developed video anomaly detectors. Hence, it is always recommended to maintain annotation consistency for all the samples of a particular dataset with the help of appropriate labeling techniques.

6.4 Lack of sufficient ground truth

The performance evaluation of the developed video anomaly detectors can be performed effectively both at qualitative and quantitative levels, provided that appropriate ground truths (both temporal and spatial ground truths) are provided. However, most of these datasets do not provide complete ground truths, i.e., both spatial and temporal annotations for all the test samples as mentioned in Tables 2 and 3. However, few datasets such as UMN [46, 47, 87], Subway [41, 80, 81], and Anomalous Behaviour [78, 79] provide only temporal ground truths, which helps in performance evaluation at frame-level only. Further, few datasets such as UCSD Pedestrian [31], CUHK Avenue [42], Boss [83], LV [49], ShanghaiTech Campus [50], Street Scene [1] and LAD [51] offer both spatial and temporal ground truths, facilitating the performance evaluation both at pixel as well as frame levels. Therefore, there is a requirement of the video anomaly datasets having both spatial and temporal annotations for all the testing samples.

6.5 Lack of good big datasets

Only a few datasets such as LV [49], ShanghaiTech Campus [50], UCF crime [7], Street Scene [1], and LAD [51] are suitable for developing video anomaly detectors using deep learning techniques. This is because data-driven approaches such as deep learning-based modeling require large datasets. However, deep learning-based methods require not only large datasets but also good quality datasets. Here, good quality datasets means datasets having annotation consistency, redundant free, clean (noise-free) high resolution, and suitable ground truth [89]. Hence, there is a requirement for good and big datasets for developing efficient video anomaly detectors.

6.6 Lack of wide diversity in the scenarios

Most of the available datasets such as Subway [41, 80, 81], UCSD Pedestrian [31], CUHK Avenue [42], UMN [46, 47, 87], LV [49], Anomalous Behavior [78], etc., cover only a limited number of scenarios in the range of one to ten. However, the video anomalies may have high variance (or a different variety of cases) within the positive samples for a limited number of available training data samples [90]. Hence, there is a need for creating video anomaly datasets that cover a wide range of anomalous activities in diverse scenarios.

7 Understanding and transforming the datasets

It is always advisable to be familiar with the dataset by understanding its inherent patterns with the help of data exploration, subsequently transforming the data samples to suitable forms with the help of data preparation. An excellent and in-depth understanding and transformation of the datasets help in better model development. Particularly, a good knowledge of the video anomaly datasets helps in selecting the appropriate modeling strategies to develop a better video anomaly detector.

7.1 Data exploration

Data exploration or Exploratory Data Analysis (EDA) is the process of understanding and identifying the interesting as well as unknown trends of the datasets by exploring graphical representations without strong dependence on the preconceived assumptions and models [91, 92]. In other words, EDA is the process of engineering the data to make them machine learning-friendly with the availability of lots of data [93]. Usually, the complexity of the EDA increases with the increase in the dimensionality of the input data. Hence, performing EDA on the high dimensional and unstructured data such as image, video, and audio datasets is slightly different from numerical data. In this regard, various types of dimensionality techniques such as Principal Component Analysis (PCA) [94–96], Multidimensional Scaling (MDS) [97], and t-distributed Stochastic Neighbor Embedding (t-SNE) [98] are found to be helpful. Data after EDA is easy to interpret as compared to raw data in shorter duration. Here, EDA is mainly intended to investigate the essential quantitative and qualitative parameters of video anomaly datasets visually [99]. Few dedicated tools or techniques [100–103] have been developed to facilitate query-based Explanatory Video Analysis (EVA). An EVA is performed for the five important public datasets of the video anomaly detection such as Subway [41], UCSD Pedestrian [31], CUHK Avenue [42], and Street Scene [1] for single scene formulation as shown in Table 6. Similarly, another EVA is performed for the five important public datasets of the video anomaly detection such as UMN [46, 47], LV [49], UCF Crime [7], ShanghaiTech Campus [50], and LAD [51] for multiple scene formulation as shown in Table 7. Particularly, these ten datasets are selected for the EVA due to the high trend in the recent citation frequency, diversity in scenarios, good quality, and quantity of the available datasets.

7.2 Data preparation

Data preparation is the process of cleansing and transforming the raw data prior to the execution of the main task of the model [104, 105]. It is the initial and essential step in the development of any data analytics application. In the case of video/image datasets, RGB to gray-scale conversion, the data preparation involves image resizing, video trimming, pixel scaling, reformatting data to a common format (standardizing data), ensuring annotation consistency, making corrections to the data, and combination of datasets to enrich data. Generally, data preparation should be performed consistently across all the data samples of the concerned datasets used for the model development. However, image /video augmentation is usually applied to the training samples only and not to the validation as well as testing samples. Data preparation helps in removing or minimizing the inherent bias (caused by

Table 6 Sample frames of the key single scene video anomaly datasets

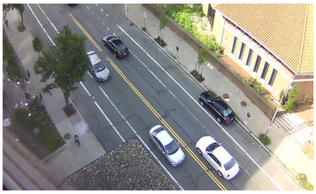
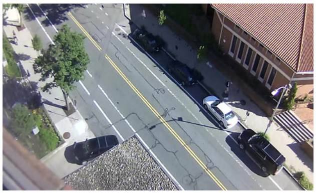
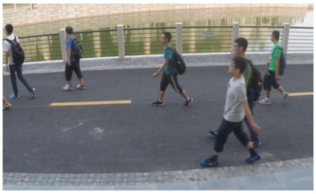
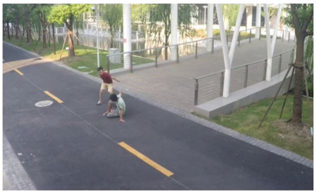
Dataset	Normal frame	Anomalous frame
Subway Entrance [41]	Turnstile crossing 	Avoiding turnstiles
Subway Exit [41]	Empty turnstile 	Wrong direction
UCSD Ped1 [31]	Walking 	Bicycle, Vehicle
UCSD Ped2 [31]	Walking 	Bicycle, Skater
CUHK Avenue [42]	Queue 	Throwing papers
Street Scene [1]	Orderly traffic 	Biker on sidewalk 

Table 7 Sample frames of the key multiple scene video anomaly datasets

Dataset	Normal frame	Anomalous frame
UMN [47, 87]	Normal crowd	Panic crowd
LV [49]	Normal market	Fighting
UCF-Crime [7]	Traffic	Explosion
ShanghaiTech Campus [50]	Pedestrians 	Brawling 
LAD [51]	Normal fire	House on fire

the poor quality of the data samples) of the datasets. For an example, video anomaly detection is preceded by the data preparation stages such as RGB to gray scale conversion and normalization [39, 40].

8 AI approaches to develop efficient video anomaly detection and localization models by solving issues related to video anomaly datasets

An AI system is built on data and codes (models or algorithms) [89]. The development cycle of an AI-based system mainly comprises four steps: defining the scope of the project, collecting and preparing data, training the model, and deploying the trained model in production. Further, there are two types of approaches, such as model-centric approach and data-centric approach, for developing the deep learning models [89]. Specifically, these two types of approaches are also suitable for developing deep learning-based video anomaly detection models. Further, a taxonomy of AI approaches to develop efficient video anomaly detection and localization models by solving issues related to video anomaly datasets is presented in Fig. 1.

8.1 Model-centric approaches

In this approach, the model (code or algorithm) is systematically improved to get the desired accuracy from the developed models [89]. In other words, the model-centric approach aims to enhance the model performances with the existing dataset by enhancing the model archi-

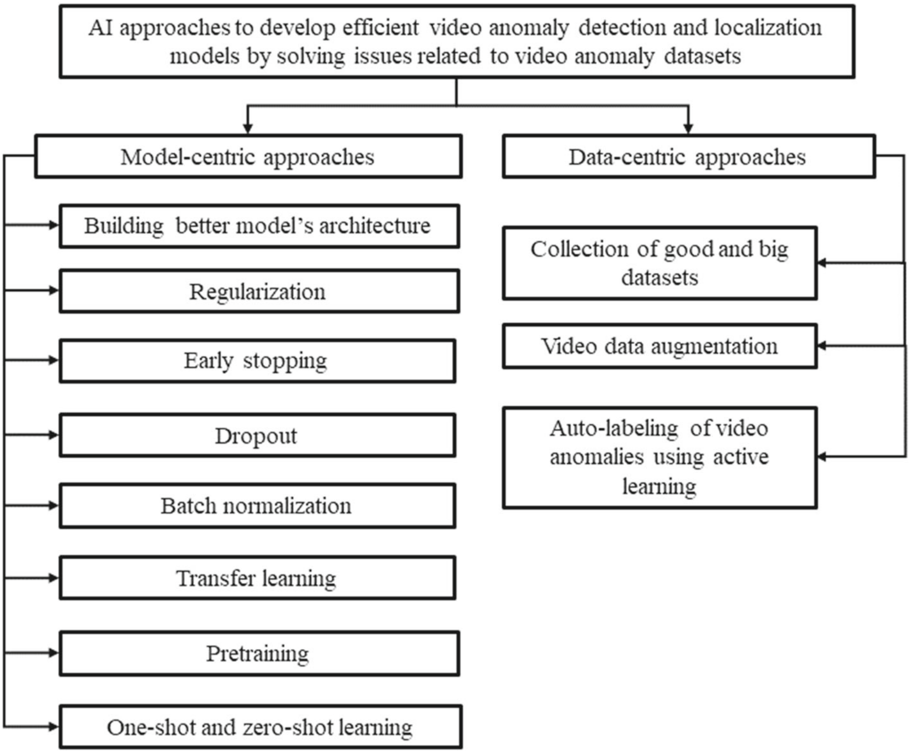

AI approaches to develop efficient video anomaly detection and localization models by solving issues related to video anomaly datasets
Model-centric approaches
Data-centric approaches
Building better model's architecture
Regularization
Early stopping
Dropout
Batch normalization
Transfer learning
Pretraining
One-shot and zero-shot learning
Collection of good and big datasets
Video data augmentation
Auto-labeling of video anomalies using active learning

Fig. 1 Taxonomy of AI approaches to develop efficient video anomaly detection and localization models by solving issues related to video anomaly datasets

tecture. So, consistently using a high-quality model throughout the development cycle is paramount. Various tools and techniques such as hyperparameter tuning, efficient and latest models, transfer learning, etc., are used to improve the model's performance. Here, development is carried out by improving the code iteratively while keeping the data fixed. Generally, in this approach, data is collected as much as possible within the project's scope, and focus is given to developing a good enough model to deal with the noise in the data. Most of the AI-based researches has been performed using a model-centric approach. In many applications, the model-centric approach helps to attain good performance with complex and deeper models, resulting in a subsequent increase in maintenance and computational cost.

Few important model-centric approach-based video anomaly detection methods are deep autoencoder with an autoregressive estimation network [106], memory-augmented autoencoder [107], object-centric Auto-encoders [108], spatiotemporal graph autoencoder [109], Message-Passing Encoder-Decoder Recurrent Neural Networkn [110], self-trained deep ordinal regression [111], variational autoencoder [112], autoencoder with memory models for learning diverse and discriminative normal patterns [113], end-to-end adversarially Learned One-Class Classifier [114], dual stream variational autocendoer [115], GAN-based Stacked variational autoencoder [116], noise-modulated GAN [117], residual-spatiotemporal translation network [118], and graph convolutional network [119]. However, there are lots of scope to improve the existing DNN architecture or develop novel DNN architectures for the video anomaly detection.

Further, few important model-centric approaches used to address the various issues related to the video anomaly datasets are outlined as follows.

8.1.1 Building better model's architecture

More robust and abstract features should be used in training for improving the model generalization ability. This can be achieved by modifying the architecture of the deep learning models. Based on this approach, significant number of progressively complex and efficient architectures of deep learning models such as AlexNet [120], VGG Net [121], ResNet [122], Inception-V3 [123], DenseNet [124], CondenseNet [125], BubbleNet [126],Convolutional Spatiotemporal Auto-Encoder (ConvSTAE) [11], etc., have been proposed. Mostly, the idea behind the development is going for the deeper and thinner networks. However, theoretically one can go to higher depths by using more and more layers [127]. But, practically, the training of the highly deeper network is not feasible due to insufficient strength of the activation signal at the higher nodes and the requirement of excessive-high computational cost. Hence, there is always a trade-off between the feasible depth of the Deep Neural Networks (DNN) and desired performance. Recently, few important research works based on enhanced architectures such as sparse denoising autoencoders [128], TransAnomaly using video vision transformer [129], SmithNet using motion-texture coherence [130], generative adversarial network using self-attention [131], faster RCNN using deep reinforcement learning [132], non-local U-Net [133], combination of GAN and frame prediction [134], etc., have been reported for the video anomaly detection and localization.

8.1.2 Regularization

A DNN is said to be suffering from over-fitting when it performs well on the training dataset and significantly worse on the testing (unseen) data of the same domain or application [127, 135]. Mostly, over-fitting is caused by the noise present in the training datasets as the model has learned it as an underlying pattern of the training data. Unfortunately, these noises are not the same for all the datasets of the same domain. Further, the DNNs may be over-fitted or under-fitted when the datasets are small in size or imbalanced. In practice, complex DNN (having more layers and neurons) is more susceptible to over-fitting than shallow neural networks. Regularization is one type of functional solution that performs small changes in the learning algorithm, improving the model generalization ability over the unseen data by reducing the variance. There are different types of regulation techniques such as L1-regularization (helps in decaying the weights to zero) and L2-regularization (helps in decaying the weights towards zero, i.e., not exactly zero).

8.1.3 Early stopping

In contrast to over-fitting, under-fitting occurs when the model fails to capture the variability of the data [136]. Here, the model possesses almost no predictive power due to the lack of proper mapping of the training data [137]. There are various penalty methods such as structural risk minimization, generalization cross-validation, etc., to avoid the problems of the over-fitting and under-fitting [136, 138]. Another approach to address this issue is by using early stopping criteria based on the results obtained for the training, validation, and testing data. When validation error starts to increase after the initial decrease, then the training is stopped to avoid over or under-fitting [136].

8.1.4 Dropout

Generally, large DNN suffers from sluggishness and over-fitting during testing as the network has to combine predictions from the various large neural networks. This problem can be addressed by using dropout, a technique to randomly discard the neurons with corresponding weights from the network during training to prevent too much co-adaption [139]. In other words, the process of making zeros out the activation strengths of the randomly selected neurons during training forces the DNN to learn more robust features instead of relying on the predictive ability resulted from the small datasets [127, 139]. Subsequently, dropout is incorporated in CNNs in the form of spatial dropout to discard the feature maps instead of single neurons [140]. The dropout rate is a hyperparameter that needs proper tuning corresponding to training datasets. Most of the reported deep learning-based video anomaly detection techniques [141, 142] use dropout to prevent the over-fitting of the DNN.

8.1.5 Batch normalization

Batch normalization is a type of regularization technique used to the set of activation values in a layer by subtracting the batch mean value from individual activation and subsequently dividing by the batch standard deviation [127, 143]. It is helpful for the preprocessing of the video sequences at the pixel level.

8.1.6 Transfer learning

Generally, the efficiency of the machine learning models mostly lies on a common assumption that both train and test data belong to the same feature space as well as the same distribution [144]. Whenever the distribution changes, it is advisable to learn the model from the scratch. However, in most real-world applications, it is not feasible or too expensive to collect and prepare the training data required to rebuild the model from scratch [145]. In this scenario, transfer learning or knowledge transfer helps the developer to address the problems associated with small datasets. Transfer learning is the process of utilizing the weights of a successfully trained model on a big dataset as the initial weights for another new problem of the related field [127]. This is very much helpful in image or video processing as many datasets share better learned low-level spatial characteristics. However, transferability is also negatively affected due to the use of higher neurons in the primary task and optimization difficulties associated with the splitting of the co-adapted neurons [146]. The effective use of transfer learning is a significant development consideration. Various deep learning-based video anomaly techniques [7, 147, 148] used transfer learning for robust and efficient feature extraction.

8.1.7 Pretraining

Though pretraining is conceptually similar to the transfer learning, they are not the same. In transfer learning, the weights along with the network architecture have to be transferred to the target model [149]. But, in the case of pretraining, the new network architecture meant for the target application is initialized with the weights of the source model trained on big datasets, i.e., pretraining provides some flexibility in the network architecture design [127].

8.1.8 One-shot and zero-shot learning

One-shot and Zero-shot learning [150, 151] are very much useful for developing the deep learning models using very limited training data [127]. In One-shot learning, the classification model is trained from very limited training datasets comprising one image (or few) per class. One-shot learning is very much helpful for the facial recognition tasks [152, 153]. Further, zero-shot learning is used to develop the deep learning-based classification models when there is an extreme scarcity of training data, i.e., very few training samples or sometimes no labeled data. Zero-shot learning uses descriptive attributes using input and output vector embedding such as Word2Vec [154] or GloVe [155] to classify the image samples [127]. Recently, a one-shot-based Siamese 3D CNN [156] has been proposed for video anomaly detection. Further, zero-shot learning-based crowd behavior analysis [157] and outlier detection [158] are presented. However, further investigation is needed on the one-shot and zero-shot learning-based video anomaly detection.

8.2 Data-centric approaches

In this approach, data are systematically improved to get the desired accuracy from the developed models [89]. In other words, the data-centric approach aims to enhance the model performances with the existing dataset by enhancing the data. So, consistently using high-quality data throughout the development cycle is of paramount importance. Various tools and techniques are used to improve the qualities such as annotation consistency, noise-free, coverage over important cases, etc. Here, development is carried out by enhancing the data iteratively while keeping the code fixed. The data can be systematically improved by changing either the data (input) or corresponding labels or both in an iterative manner to achieve the desired performance. In this approach, the focus is also shifted from big data to good data. In other words, only big data is not sufficient enough to attain better performance. Instead, good (high-quality) data is required throughout the development cycle. Good data should have the following important characteristics. The data is defined consistently (definitions of annotations are unambiguous), has a high value of uniqueness (free from duplicates), and data cover all the important cases (good coverage of the input dataset for the desired application). The data has improved significantly by getting timely feedback from the production data whose distribution covers data drift and concept drift. The data should be sized appropriately while ensuring no bias is present. In many practical scenarios, the data-centric approach helps to attain good performance with simple and lightweight models resulting in a subsequent reduction in maintenance and computational cost. However, a significantly less percentage of researches has been performed using a data-centric approach. Mostly, it is due to the tedious and costly process of video anomaly dataset collection, preprocessing, and preparation.

Few important data-centric approach-based video anomaly detection methods are data augmentation in temporal domain [39], and good video anomaly datasets such as Live Videos (LV) [49], UCF crime [7], ShanghaiTech Campus [50], Large-scale Anomaly Detection (LAD) [51], A Day on Campus (ADOC) [67], learning with annotations of different degrees [159], etc. However, there are a lot of scopes to improve the existing datasets, develop better datasets and efficient data augmentation techniques for video anomaly detection.

Data-centric approaches try to solve the various issues such as over-fitting, under-fitting, class-imbalance, etc., from the root of the problems by improving the training datasets. Few important data-centric approaches used to address the various issues related to the video anomaly datasets are outlined as follows.

8.2.1 Collection of good and big datasets

The performance of the deep learning-based models can be improved significantly by using good and big datasets. Here, the qualifier ‘good’ signifies the high level of annotation consistency (error-free annotations) and data uniqueness (free from unwanted duplication of data samples) covering various scenarios. Similarly, the quantifier ‘big’ denotes the size of the video anomaly datasets. Hence, collecting good and big datasets is one of the most challenging and primary assignments in the deep learning development cycle. Higher quality and quantity of the video anomaly datasets automatically result in better model performance. Recently, few good and bog datasets such as such as LV [49], UCF crime [7], ShanghaiTech Campus [50], LAD [51], ADOC [67], UBnormal [52], etc. have been developed to achieve better performance of the video anomaly detectors.

8.2.2 Video data augmentation

Image data augmentation techniques have been studied in-depth, whereas very few studies are available for video data augmentation. In the case of an image, only spatial augmentation is feasible. However, in the case of videos, both spatial and temporal augmentations are possible. Temporal augmentation can be applied mainly in two ways, such as cropping random sequences from consecutive frames and subsampling of the video frames at different frequencies using various stride sizes. For example, all the video clips comprising of the frames 1, 2, 3, 4, 5, 6, 7, 8, ... for stride = 1; frames 1, 3, 5, 7, 9, ... for stride = 2, frames 1, 4, 7, 10, ... for stride = 3, and so on, will have the same label [39, 40]. Subsequently, more video clips are generated from the single video clip to increase the volume of the data. The spatial augmentation is the same as applying it to the images, where spatial augmentation is applied to all the frames (or images) of the complete video clip instead of an individual image. It is always advisable to use these techniques for video augmentation without simply duplicating the same video clips to achieve data uniqueness. Particularly for video data augmentation, temporal augmentation can be independently applied. However, spatial augmentation must be applied for all the frames of the video clips to maintain temporal coherence. These types of video data augmentation are effective in video classification, video anomaly detection, etc.

Basically, the spatial augmentation in the video can use most of the image data augmentation techniques. A comprehensive survey of image data augmentation techniques can be found in [127]. The important and widely used image data augmentation approaches can be broadly classified into the three categories such as approaches based on basic image manipulations, deep learning, and meta-learning.

The approaches based on basic image manipulations aim to increase the quantity and quality of the image data using various geometric transformations (flipping, cropping, rotation, and translation), color space transformations (manipulation in color space and noise injection), kernel filters, random erasing, and mixing images [127]. All these augmentation techniques are applied to the images at the input space. However, safer data augmentation techniques should ensure the level of preservation for the annotated data.

Recently, approaches based on deep learning have been successfully applied for image data augmentation by exploring the potential of the DNNs. The most widely used techniques in deep learning-based image data augmentation approaches are feature space augmentation [160], neural style transfer [161, 162], adversarial training [163], Generative Adversarial Network (GAN)-based data augmentation [164], and so on. However, these types of image augmentation techniques always require high computational complexity.

Further, another set of approaches based on meta-learning have been explored for image data augmentations. Meta-learning is used in deep learning for optimizing neural networks with the help of neural networks [127]. The most widely used techniques in meta-learning approaches are neural augmentation, auto-augment, and smart augmentations. Neural augmentation is a strategy to meta-learn a neural style transfer strategy [165]. It is suggested that the best strategy will be a combination of traditional augmentations and neural augmentations. Smart augmentation is another technique for meta-learning augmentations where the combined images are exclusively derived from the learned parameters of a pre-trained CNN instead of a neural style transfer algorithm as used in neural augmentation [166]. However, auto-augment is a different meta-learning approach as compared to neural augmentation and smart augmentation as it is based on the reinforcement learning concept [167]. Auto-augment tries automatically to find an optimal augmentation policy from a constrained set of geometric transformations having the miscellaneous level of distortions [168]. However, meta-learning is a relatively new, time-consuming, and complex approach. Hence, there is a requirement for detailed investigation for meta-learning.

Moreover, in the case of video, occlusion degrades the effectiveness of the spatial augmentation techniques when applied to the video frames directly. This is because the moving foreground object may act as an occlusion for the background region intended for the augmentation [169]. Further, the complexity increases with the camera movement. Hence, multiple camera feeds can be used for image-based rendering for video augmentation using perspective camera model [170]. Hence, the labeling should be occlusion-aware. In this regard, a video data augmentation technique based on occlusion-aware and uncertainty-enabled label propagation algorithm [171] is proposed for semantic segmentation [172]. Subsequently, GAN-based video data augmentation technique using dynamic image (motion information of the video is compressed into a still image by removing the interference event like the background) is proposed [173]. Recently, a novel video augmentation technique known as

VideoMix is proposed to improve the performance of the video classifier significantly [174]. Here, a new training video is created by inserting a video cuboid into another training video while mixing the ground truth labels proportionally to the number of voxels of each video. Moreover, there is little progress in the video data augmentation techniques delicately suitable for unsupervised tasks such as video anomaly detection. Recently, a GAN-based data augmentation technique using doping is proposed for unsupervised anomaly detection [175]. Here, the objective is to increase the size of the training datasets by oversampling the rare normal samples. Later, another technique of generating the abnormal data by transforming the normal data to address the data-imbalance problem is proposed [176]. There is ample scope to increase the quality of the video anomaly datasets to address various issues like the high false alarm, data imbalance, and so on with the help of efficient video data augmentation techniques.

Generally, all the data augmentation techniques mentioned above are applied to training datasets only. However, recently it has been reported that performing data augmentation during testing, i.e., test-time augmentation, increases the effectiveness of the model [127]. A test-time augmentation is a data-centric approach in the data space similar to ensemble learning in the model-centric approach. Further, test-time augmentation is a process of data distillation that can describe the effectiveness of ensemble predictions to get a better representation of the images/frames [177]. When a model provides better predictions (having low variance) across the augmentations, i.e., both for the training and test-time augmentations, then the particular model is said to be a robust model [127]. However, it is tough to aggregate all the predictions over the augmented test frames for video anomaly detection. Further, test-time augmentation techniques are somehow good for supervised learning tasks. The test-time augmentation increases the inference complexity and hence, the cost of the model.

8.2.3 Auto-labeling of video anomalies using active-learning

Manual Identification and labeling of the anomalous video segments are time-consuming, laborious, erroneous, and costlier jobs [35]. This can be overcome with the help of auto-labeling of video anomalies using active learning [36]. The video anomaly detector is initially trained with the available datasets using a model-centric approach and deployed in the targeted scenarios for real-time field trials. The initial video anomaly detection model detects the video anomalies and labels them automatically. Then these videos are stored in a temporary location and available for human domain expert verification using the developed graphical user interface. The domain expert has to either approve or reject the decision of the video anomaly detector. When the annotators approve the decisions, the machine-generated labels are preserved. However, when the annotator rejects the decisions, there will be provision for correcting the labels by the annotators. After corrections, the datasets can be added to the final database. In this way, a new and good dataset will be available for retraining the model over time. Then, the model can be retrained and deployed in the field. This cycle can be repeated few times for adding the changes to the model due to new changes occurring in the scenarios. In this way, a robust and efficient deep learning-based video anomaly detector model can be developed. This approach helps in speeding up the initial labeling step and is more suitable for online applications. However, the model requires continuous human intervention.

It is not always possible for practical scenarios to get the best performance, either using model-centric or data-centric approaches independently. Therefore, it is always better to use a hybrid approach, i.e., a systematic combination of both model-centric and data-centric approaches for AI systems to get better performances. Initially, the AI system should be developed by using the available data and model to get a baseline performance. Then, a data-centric approach should be applied to enhance the performance, followed by a model-centric approach. This cycle of development using data-centric and model-centric approaches can also be performed iteratively to get the best performance from the developed model corresponding to a particular application. In summary, one has to give focus not only on a code but also on data to develop an efficient AI system.

9 Emerging trends in detection and localization of video anomalies

The research and development trends in video anomaly detection are fast-moving due to the availability of massive computational resources, efficient AI techniques, good and big video anomaly datasets. Hence, key emerging techniques and their alignments with the availability of the video anomaly datasets are presented as follows.

9.1 Problem formulation trends

Nowadays, multiple scene video anomaly datasets are in trend due to the availability of the good and big video anomaly datasets generated from the readily available multiple video surveillance cameras mounted at various geographical locations. Hence, video anomaly detection and location using MSF is gradually taking the lead as compared to SSF-based video anomaly detection. Hence, the video anomaly detectors are expected to be more generalized without compromising on the anomaly accuracy detection rate.

9.2 Modeling trends

Usually, detection and localization of the video anomalies are modeled using representation, predictive, one-class, and generative models. Recently, due to the availability of good and big video anomaly datasets and massive computational facilities to process them, deep hybrid models (a combination of more than one modeling technique) for achieving better anomaly detection accuracy with competitive processing speed.

9.3 Trend in training and learning frameworks

Inherently, video anomaly detection is an unsupervised learning problem. Further, video anomaly datasets are inherently imbalanced. Labeling all the normal samples at the frame level is a tedious and erroneous job. However, weakly labeled (folder-level annotated) normal video samples are readily available in both good quality and quantity. Hence, weakly-supervised learning is recently in trend for video anomaly detection. Video anomaly detection based on a weakly supervised learning approach uses only folder-level annotated normal video samples during training. However, the weakly supervised trained video anomaly detector is capable of localizing the video anomalies in the spatiotemporal domain.

9.4 Evolution of evaluation criteria

The evaluation of video anomaly detection and localization is commonly carried out utilizing three criteria, namely Frame-Level (FL), Pixel-Level (PL), and Dual-Pixel-Level (DPL) [22, 44, 65, 86]. However, rapid improvements in video anomaly detection and localization methods driven by the availability of large-scale video anomaly datasets and massive computational infrastructure also ensure the development of effective and robust evaluation strategies. New evaluation criteria such as Object-Level (OL) [4, 86] and Region-Level (RL) [1]. based on Intersection Over Union (IOU) are proposed. Similarly, the RL criterion, namely Region-Based Detection Rate (RBDR), can be used to measure the effectiveness of the VAD methods employed. The recent trend in evaluation criteria suggests that one must evaluate the performance of the proposed video anomaly detection and localization methods using primary criteria such as FL, PL, and DPL criteria with advanced criteria such as RL as well as OL to ensure the overall effectiveness of the deployed model. However, there is ample scope to develop lightweight, effective, and robust evaluation criteria with appropriate performance metrics.

9.5 Current trajectory in computational resource requirement

There is an increasing trend in the requirement for massive computational platforms with high-end GPUs, processors, RAM, and storage for training the DL models with large-scale video anomaly datasets to detect and localize the video anomalies.

9.6 Targeting better trade-off between detection accuracy and inference speed

Nowadays, IVSS is in high demand to detect video anomalies in real-time with higher accuracy and lower latency to meet the deadline of latency-sensitive applications [178]. Hence, there is a high requirement for lightweight, robust, and efficient deep hybrid models to detect and localize the video anomalies at the edge only. Hence, researchers are always targeting to achieve a better trade-off between anomaly detection accuracy and inference speed of the deployed model corresponding to the desired application.

9.7 Progressive shift from mono-modal VAD to multi-modal VAD

Multi-modal video anomaly detection is a sophisticated methodology employed for detecting the presence of anomalies within video data. This strategy entails the amalgamation of several sources or modalities of information to enhance the accuracy and effectiveness of the detection process [179]. The concept of multi-modal anomaly detection involves the integration of many data modalities, including visual, audio, text, and sensor data, with the aim of enhancing the accuracy and resilience of anomaly detection within video streams [180]. Hence, there is a progressive shift from mono-modal VAD to multi-modal VAD, which is at the barely minimum early stage of the research due to unavailability of bench marked public multi-modal video anomaly datasets.

10 Applications of detection and localization of video anomalies

Video anomaly detection and localization has a wide range of potential applications, i.e., most of the IVSS can use video anomaly detection and localization to enhance their performances. Following are the few crucial applications of video anomaly detection and localization across diverse applications fields catalyzed by the availability of large-scale video anomaly datasets.

10.1 Smart city applications

Smart city is a recently emerged concept to improve the quality of life of the urban citizens by converging information and communication technologies with the Internet of Things-enabled physical infrastructure of the city [181]. Ensuring the safety of lives and assets at public places is one of the important components of the smart city. Hence, the IVSS can use video anomaly detectors to detect and localize various anomalous activities using live video streams. For example, various anomalous activities such as abuse, burglary, explosion, fighting, riots, vandalism, and so on can be detected in time using video anomaly detectors [182]. However, video anomaly detectors must be trained with sufficiently good and big datasets corresponding to these anomalous activities and application scenarios.

10.2 Intelligent transport systems

Video anomalies such as abnormal activities/events inside the vehicle and on the roads can be detected from the video feeds using appropriate models for video anomaly detection. Various on-road anomalies such as wrong U-turn, overtaking, road accident, etc., can be detected in real time [183–185]. Subsequently, emergency service providers such as hospitals, police stations, etc., can be notified promptly to prevent further damage or loss of lives and assets.

10.3 Defense applications

Video anomaly detectors can be used for various defense applications such as border surveillance, intrusion detection, abnormal activity detection happening near-critical defense establishments, etc., from the video feed either from fixed IP cameras or from moving cameras (e.g., cameras mounted on drones). One major constraint is the unavailability of the proper datasets for these sensitive applications.

10.4 Online education

Recently, the traditional offline mode of teaching is rapidly shifting to the online mode of teaching due to the tremendous enhancements of the ICT technologies. Particularly, since the outbreak of COVID-19, online modes of education, including teaching, learning, and evaluation, have been widely used [186]. Video anomaly detection techniques can be used to automatically detect the suspicious or abnormal behaviors such as cheating, malpractices, etc., during the proctored examination over the existing ICT infrastructure [187]. Though few works [188, 189] have been reported in this direction, there are lots of scopes to propose more robust and intelligent techniques to detect anomalous behaviors during the proctored examination.

10.5 Smart home

Now a days, many people are installing video surveillance systems at their residential premises to ensure comprehensive security of their lives and assets [186]. However, video anomaly detection algorithms can run on top of the video surveillance system to automatically detect abnormal activities such as intrusion [190], loitering [191], thefts [192], etc., in real-time so that appropriate preventive actions can be initiated.

10.6 Public health management

The objective of public health management is to improve the quality of human life through the appropriate prevention and treatment of infectious diseases. For example, video anomaly detection can be used to detect anomalous behaviors such as violation of social distancing, no proper use of face masks, etc., for preventing further transmission of infectious diseases like COVID-19. Though few works [186, 193–197] have been reported in this regard, there is ample scope to propose further lightweight and robust deep learning-based models for the automatic and timely detection of anomalous behaviors.

10.7 Industrial applications

Accurate and timely detection of the video anomalies in industrial premises will help the early detection of potential breakdowns and proactive maintenance activities [186, 198]. Various video anomalies related to industrial applications are production breakdown, discontinuity of production process, objects/accessories at the wrong places, etc. Industrial applications could be improved significantly with the help of deep learning-driven video anomaly detection, Digital Twin technologies [199] , virtual and augmented realities for better productivity [200, 201].

11 Research challenges and future directions

Important research challenges associated with video anomaly detection due to the lack of big and good datasets may be outlined as follows.

11.1 Difficulty in collecting and generating video anomaly datasets covering all the environmental conditions

Collection, annotation, and preprocessing of video datasets is a tedious, time-consuming, and costly job. The deep learning model performs better only in a similar scenario with the training samples. So, to make the video anomaly detector that will efficiently work in the day, night, sunny hours, rainy hours, camera shaking due to high wind speed, etc., requires video datasets covering all these scenarios. Practically, it is not always feasible to collect the video datasets covering all these scenarios. However, this problem can be minimized significantly by choosing appropriate video data augmentation techniques to generate similar artificial videos corresponding to these environmental conditions from the available training datasets.

11.2 Difficulty in developing end-to-end trainable deep learning-based video anomaly detector

Broadly, video anomaly detection is the process of detecting and localizing the abnormal events or unexpected spatiotemporal patterns. However, in practical scenarios, there is no clear boundary between the normal and anomalous events due to the inherent equivocal natures of the video anomalies [186]. Hence, perfect labelling of the video anomalies in the spatiotemporal dimensions is not always possible [9, 10, 22, 202]. So, insufficient labeled data prevents the development of complete end-to-end trainable deep learning-based video anomaly detector [9, 10, 22, 202]. However, good and big datasets such as ShanghaiTech Campus [50], Street Scene [1], and LAD [51] are published recently due to rapid advancement in video surveillance technology. Hence, the development of end-to-end trainable deep learning-based video anomaly detectors is feasible for the specific context.

11.3 Difficulty in applying supervised learning techniques

Anomalies are inherently rare occurring events, and hence, the class imbalance problem is always associated with the video anomaly detection [22, 203, 204]. Due to imbalanced datasets, the models are biased towards the majority class predictions (normal events), leading to an increase in miss rate for video anomaly detection. Subsequently, supervised learning-based video anomaly detection is not a reliable modeling approach. However, data augmentation is a data-level solution that helps in alleviating the data imbalance problem [127]. Hence, supervised learning-based video anomaly detection is just started due to the availability of good and big datasets like LAD [51] and UCF Crime [7].

11.4 Difficulty in developing universal video anomaly detector

Video anomalies are always context-dependent [12, 22, 205]. A video anomaly detector meant for a classroom scenario may not be suitable for the market scenario. So, the development of a universal video anomaly detector is not possible, i.e., the video anomaly detector models are not adaptable [186]. However, there is scope to develop a video anomaly detector for a wider range of applications that can cover multiple scenes using good and big datasets covering multiple appropriate scenarios.

11.5 Selection of appropriate training strategy with the available video datasets

Selection of appropriate learning mechanisms such as supervised, semi-supervised, weakly supervised, and unsupervised is carried out depending on the available dataset and targeted applications. Recently, a novel learning mechanism, i.e., curriculum learning [206], is proposed to select the training data in a systematic way that increases the model performances as compared to the random selection [127, 207–209]. Further, curriculum learning is applicable for all the deep learning models (including models trained with limited data). The curriculum learning helps to understand the underlying patterns in the training data by plotting the training accuracy over time for various initial training subsets. This understanding allows the developer to speed up the training process [127]. Generally, data augmentation yields massive training data resulted from applied augmentation techniques, and hence, there is a high chance to get a training subset that is responsible for faster and more accurate training. It is suggested to train the model with original datasets initially, and subsequently, the training should be completed with the datasets comprising original as well as augmented data points. However, there is no clear consensus, and hence, no standard practice regarding the way to perform curriculum learning [127] has been developed. Further, it is complicated to apply curriculum learning in the case of video data as complexity increases significantly. As per our survey, no work related to video anomaly detection has used curriculum learning yet. However, it may be used in video anomaly detection in the future due to the availability of high-end computational facilities.

11.6 Selection of appropriate resolution

Deep learning-based models require extensive computational resources to be trained using higher resolution videos/images. Therefore, most of the deep learning models downsample the resolution of the HD and 4K videos/images before processing. However, downsampling of the resolution increases the speed at the cost of accuracy [210–212]. Hence, there is always a trade-off between accuracy and speed during training using down-sampled videos/images. An ensemble model trained with both high and low resolution performs better as compared to that of individual models [127]. Hence, choosing an appropriate combination of low and high-resolution samples in the training video datasets is a challenge to get better model performance.

11.7 Difficulty in handling high voluminous data resulted from video augmentation

One of the crucial aspects of video data augmentation is to determine the size of final datasets as the video itself is highly computational hungry. If the final datasets, i.e., the combination of the original samples and augmented samples, is very large, then it may not be feasible to handle them by providing the required additional memory and computational power. When data augmentation is applied on the fly (online data augmentations) using the data transformers embedded in the data loaders, then memory storage requirement reduces at the cost of training time. In contrast, if all the augmented data are prepared using independent transformers before the training (offline data augmentations), then training time decreases at the cost of memory requirement [127]. Hence, the developers should decide the appropriate augmentation technique depending on the available resources and targeted applications. Recently, it is possible to apply video data augmentation using a massively distributed training system using offline data augmentation [213, 214]. However, it is better to use a few but effective video data augmentation techniques to achieve desire model performance in the case of video anomaly detection.

11.8 Privacy issue

Video anomaly detection is based on the video surveillance data that includes facial and behavioral information of the subjects available on the scene. However, an individual's right to privacy is breached when the video anomaly dataset is made publicly available as an open-source dataset [186]. Hence, there is a scarcity of open-source video anomaly datasets

11.9 Noise issue

Video surveillance data is readily available due to the wide use of video surveillance systems at various public places such as shopping malls, market areas, traffic junctions, homes, etc. However, manual data preparation and annotation processes are tedious and more prone to errors [186]. Hence, there is a high chance that noise will be inherently present in the video data, and eradicating the noise from the video data is not feasible. Hence, this noise will degrade the quality of the dataset, and subsequently, the model performance will be negatively affected.

11.10 Need for hybrid approach

Neither the model-centric approach nor the data-centric approach is self-sufficient in standalone mode for developing an efficient video anomaly detector. Therefore, there is a need for a hybrid approach involving both model-centric and data-centric approaches in a complementary manner to develop efficient DL models for the video anomaly detection. Recently, data-centric approaches have helped in enhancing the performance of the models developed by the model-centric approach due to the availability of large-scale video anomalies and massive computational resources.

12 Conclusions

The successful development of the efficient and robust deep learning-based models for the video anomaly detection as well as localization requires appropriate big and good video anomaly datasets. Hence, it is a basic requirement to understand various publicly available video anomaly datasets for effective modeling. In this article, a comprehensive review of publicly available video anomaly datasets has been presented. A brief on problem formulation depending on the training datasets using either SSF or MSF is presented. Most of the important and widely used video anomaly datasets are explained to explore data-centric approaches for video anomaly detection and localization. A comparative analysis of the publicly available video anomaly datasets is presented to explain both qualitative and quantitative parameters. Subsequently, various issues such as the fewness of the datasets, data imbalance, annotation inconsistency, lack of sufficient ground truth, lack of good big datasets, and lack of wide diversity in the scenarios related to available datasets have been summarized. Further, the data exploration and data preparation techniques required to understand and transform the datasets are presented. Various model-centric and data-centric approaches that are useful for addressing the model performance degradation due to the issues related to the datasets are discussed. It is recommended to use both data-centric and model-centric approaches in an interlaced manner to get the best possible model performance. Current research trends and challenges, particularly due to the unavailability of suitable video anomaly datasets and potential applications of video anomaly detection, are summarized. This article is expected to serve as a standalone reference for understanding the details of the publicly available video anomaly datasets.

Acknowledgements This work is supported by the **IMP**ACTING **R**ESEARCH **IN**NOVATION AND **T**ECHNOLOGY (IMPRINT) INDIA (Grant No.-7794/2016), an initiative of Ministry of Human Resource Development, Government of India and Ministry of Housing and Urban Affairs, Government of India.

Data availability The datasets analyzed during the current study are publicly available in the corresponding repositories, and they have been cited in the reference list.

Code Availibility Not applicable as the manuscript presents a comprehensive review of datasets for detection and localization of video anomalies with a focus on data-centric artificial intelligence-based video anomaly detection.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ramachandra B, Jones M (2020) Street scene: A new dataset and evaluation protocol for video anomaly detection. In: Proceedings of the IEEE winter conference on applications of computer vision, pp 2569–2578
2. Tripathi RK, Jalal AS, Agrawal SC (2018) Suspicious human activity recognition: a review. Artif Intell Rev 50(2):283–339
3. 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
4. Pawar K, Attar V (2019) Deep learning approaches for video-based anomalous activity detection.World Wide Web 22(2): 571–601
5. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT press, Cambridge, Massachusetts, London, England. http://www.deeplearningbook.org
6. LeCun Y, Bengio Y, Hinton G (2015) Deep learning.nature 521(7553):436–444
7. 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
8. Patil N, Biswas PK (2016) A survey of video datasets for anomaly detection in automated surveillance. In: Proceedings of the Sixth IEEE international symposium on embedded computing and system design (ISED), pp 43–48
9. Vu H, Phung D, Nguyen TD, Trevors A, Venkatesh S (2017) Energy-based models for video anomaly detection
10. Medel JR, Savakis A (2016) Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv:1612.00390
11. Nayak R, Pati UC, Kumar Das S (2020) Video anomaly detection using convolutional spatiotemporal autoencoder. In: Proceedings of the IEEE international conference on contemporary computing and applications (IC3A), pp 175–180
12. Kiran BR, Thomas DM, Parakkal R (2018) An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. J Imag 4(2):1–15
13. Sabokrou M, Fayyaz M, Fathy M, Moayed Z, Klette R (2018) Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Comp Vision Image Underst 172:88–97
14. Ko T (2011) A survey on behavior analysis in video surveillance applications. In: Lin W (ed) Video surveillance. IntechOpen, Rijeka Chap. 16. https://doi.org/10.5772/15302
15. Popoola OP, Wang K (2012) Video-based abnormal human behavior recognition-a review. IEEE Trans Syst, Man, and Cybernetics, Part C (Applications and Reviews) 42(6):865–878
16. Sodemann AA, Ross MP, Borghetti BJ (2012) A review of anomaly detection in automated surveillance. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42(6):1257–1272
17. Patrikar DR, Parate MR (2022) Anomaly detection using edge computing in video surveillance system. Int J Multimed Inf Ret 11(2):85–110
18. Mabrouk AB, Zagrouba E (2018) Abnormal behavior recognition for intelligent video surveillance systems: A review. Expert Syst Appl 91:480–491
19. Chalapathy R, Chawla S (2019) Deep learning for anomaly detection: A survey. arXiv:1901.03407
20. Kumaran SK, Dogra DP, Roy PP (2019) Anomaly detection in road traffic using visual surveillance: A survey. arXiv:1901.08292
21. Aggarwal CC (2017) An Introduction to Outlier Analysis, pp 1–34. Springer, Cham. https://doi.org/10.1007/978-3-319-47578-3sps1
22. Nayak R, Pati UC, Das SK () A comprehensive review on deep learning-based methods for video anomaly detection. Image Vis Comput 106:104078. https://doi.org/10.1016/j.imavis.2020.104078
23. Pang G, Shen C, Cao L, Hengel AVD (2021) Deep learning for anomaly detection: A review. ACM Comput Surv (CSUR) 54(2):1–38
24. Tran TM, Vu TN, Vo ND, Nguyen TV, Nguyen K (2022) Anomaly analysis in images and videos: A comprehensive review. ACM Comput Surv 55(7):1–37
25. Zhang D, Gatica-Perez D, Bengio S, McCowan I (2005) Semi-supervised adapted hmms for unusual event detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR’05), vol. 1, pp 611–618
26. Zhong H, Shi J, Visontai M (2004) Detecting unusual activity in video. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR), vol. 2, p. IEEE
27. Xiang T, Gong S (2008) Incremental and adaptive abnormal behaviour detection. Comp Vision Image Underst 111(1):59–73

28. Wang X, Ma X, Grimson WEL (2007) Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans Pattern Anal Mach Intell 31(3):539–555
29. Jacobs H (1967) To count a crowd. C Journal Rev 6(1):37
30. Kaltsa V, Briassouli A, Kompatsiaris I, Hadjileontiadis LJ, Strintzis MG (2015) Swarm intelligence for detecting interesting events in crowded environments. IEEE Transactions on Image Processing 24(7):2153–2166. https://doi.org/10.1109/TIP.2015.2409559
31. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1975–1981
32. Saligrama V, Chen Z (2012) Video anomaly detection based on local statistical aggregates. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2112–2119 IEEE
33. Leach MJ, Sparks EP, Robertson NM (2014) Contextual anomaly detection in crowded surveillance scenes. Pattern Recogn Lett 44:71–79
34. Zhu Y, Nayak NM, Roy-Chowdhury AK (2012) Context-aware activity recognition and anomaly detection in video. IEEE Journal of Selected Topics in Signal Processing 7(1):91–101
35. Varadarajan J, Subramanian R, Ahuja N, Moulin P, Odobez J-M (2017) Active online anomaly detection using dirichlet process mixture model and gaussian process classification. In: Proceedings of the IEEE winter conference on applications of computer vision (WACV), pp 615–623
36. Pimentel T, Monteiro M, Veloso A, Ziviani N (2018) Deep active learning for anomaly detection. arXiv:1805.09411
37. Liu Y, Li Z, Zhou C, Jiang Y, Sun J, Wang M, He X (2020) Generative adversarial active learning for unsupervised outlier detection. IEEE Trans Knowl Data Eng 32(8):1517–1528. https://doi.org/10.1109/TKDE.2019.2905606
38. Leyva R, Sanchez V, Li C-T (2017) Video anomaly detection with compact feature sets for online performance. IEEE Trans Image Proc 26(7):3463–3478
39. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 733–742
40. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. In: International symposium on neural networks, pp 189–196 Springer
41. 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. https://doi.org/10.1109/TPAMI.2007.70825
42. 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
43. Li W, Mahadevan V, Vasconcelos N (2014) Anomaly detection and localization in crowded scenes. IEEE Trans Pattern Anal Mach Intell 36(1):18–32. https://doi.org/10.1109/TPAMI.2013.111
44. 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 on Image Processing 26(4):1992–2004
45. Sabokrou M, Fathy M, Moayed Z, Klette R (2017) Fast and accurate detection and localization of abnormal behavior in crowded scenes. Mach Vision Appl 28(8):965–985
46. UMN:Unusual Crowd Activity Dataset of University of Minnesota. http://mha.cs.umn.edu/movies/crowdactivity-all.avi
47. Bird N, Atev S, Caramelli N, Martin R, Masoud O, Papanikolopoulos N (2006) Real time, online detection of abandoned objects in public areas. In: Proceedings of the IEEE international conference on robotics and automation (ICRA), pp 3775–3780
48. Blunsden S, Fisher R (2010) The behave video dataset: ground truthed video for multi-person behavior classification. Annals of the BMVA 4(1–12):4
49. Leyva R, Sanchez V, Li C-T (2017) The lv dataset: A realistic surveillance video dataset for abnormal event detection. In: Proceedings of the 5th international workshop on biometrics and forensics (IWBF), pp 1–6
50. 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
51. Wan B, Jiang W, Fang Y, Luo Z, Ding G (2021) Anomaly detection in video sequences: A benchmark and computational model. IET Image Processing 15(14):3454–3465
52. Acsintoae A, Florescu A, Georgescu M-I, Mare T, Sumedrea P, Ionescu RT, Khan FS, Shah M (2022) Ubnormal: New benchmark for supervised open-set video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20143–20153

53. Fisher RB (2004) The PETS04 surveillance ground-truth data sets. In: Proceedings of the 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp 1–5
54. Li J, Hospedales TM, Gong S, Xiang T (2010) Learning rare behaviours. In: Asian conference on computer vision, pp 293–307 Springer
55. Wang M, Wang X (2011) Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR), pp 3401–3408
56. Ferryman J, Shahrokni A (2009) PETS2009: Dataset and challenge. In: Proceedings of the Twelfth IEEE international workshop on performance evaluation of tracking and surveillance, pp 1–6
57. PETS: (2009) Performance Evaluation of Tracking and Surveillance (PETS) 2009 Benchmark Data provided by CVPR. http://www.cvg.reading.ac.uk/PETS2009/a.html
58. Benezeth Y, Jodoin P-M, Saligrama V, Rosenberger C (2009) Abnormal events detection based on spatio-temporal co-occurences. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2458–2465
59. Loy CC, Xiang T, Gong S (2009) Modelling multi-object activity by gaussian processes. In: British machine vision conference (BMVC), pp 1–11 Citeseer
60. Loy CC, Xiang T, Gong S (2010) Stream-based active unusual event detection. In: Proceedings of the Asian conference on computer vision, pp 161–175 Springer
61. Loy CC, Xiang T, Gong S (2011) Detecting and discriminating behavioural anomalies. Pattern Recognition 44(1):117–132
62. Loy CC, Hospedales TM, Xiang T, Gong S (2012) Stream-based joint exploration-exploitation active learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1560–1567
63. Xu K, Sun T, Jiang X (2020) Video anomaly detection and localization based on an adaptive intra-frame classification network. IEEE Trans Multimed 22(2):394–406
64. Wu P, Liu J, Shen F (2020) A deep one-class neural network for anomalous event detection in complex scenes. IEEE Trans Neural Netw Learn Syst 31(7):2609–2622
65. Zhou JT, Du J, Zhu H, Peng X, Liu Y, Goh RSM (2019) Anomalynet: An anomaly detection network for video surveillance. IEEE Trans Inf Foren Secur 14(10):2537–2550
66. Hospedales T, Gong S, Xiang T (2012) Video behaviour mining using a dynamic topic model. Int J Comput Vis 98(3):303–323
67. Pranav M, Zhenggang L, K SS (2020) A day on campus - an anomaly detection dataset for events in a single camera. In: Proceedings of the asian conference on computer vision (ACCV)
68. Hosmer P (2007) i-lids bag and vehicle detection challenge. In: IEEE international conference on advanced video and signal based surveillance
69. Javan Roshtkhari M, Levine MD (2013) Online dominant and anomalous behavior detection in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2611–2618
70. Ali S, Shah M (2007) A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–6
71. UCF: Abnormal Crowd Behavior Detection Using Social Force Model by UCF Center for Reserch in Computer Vision. https://www.crcv.ucf.edu/projects/Abnormal_Crowd/
72. Mehran R, Oyama A, Shah M (2009) Abnormal crowd behavior detection using social force model. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 935–942
73. Kuettel D, Breitenstein MD, Van Gool L, Ferrari V (2010) What's going on? discovering spatio-temporal dependencies in dynamic scenes. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 1951–1958
74. Jouneau E, Carincotte C (2011) Particle-based tracking model for automatic anomaly detection. In: Proceedings of the 18th IEEE international conference on image processing, pp 513–516
75. Hospedales T, Gong S, Xiang T (2009) A markov clustering topic model for mining behaviour in video. In: Proceedings of the IEEE 12th international conference on computer vision, pp 1165–1172
76. Jiang F, Wu Y, Katsaggelos AK (2008) Abnormal event detection based on trajectory clustering by 2-depth greedy search. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, pp 2129–2132
77. Oh S, Hoogs A (2011) Anomaly detection from videos under sparse data and partial observations
78. Zaharescu A, Wildes RP (2018) Anomalous Behavior Data Set. http://vision.eecs.yorku.ca/research/anomalous-behaviour-data/
79. Zaharescu A, Wildes R (2010) Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing. In: Proceedings of the european conference on computer vision, pp 563–576 Springer

80. Derpanis KG, Wildes RP (2010) Dynamic texture recognition based on distributions of spacetime oriented structure. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, pp 191–198
81. Derpanis KG, Gryn JM (2005) Three-dimensional nth derivative of gaussian separable steerable filters. In: Proceedings of the IEEE international conference on image processing 2005, vol 3, p 553
82. Hassner T, Itcher Y, Kliper-Gross O (2012) Violent flows: Real-time detection of violent crowd behavior. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops, pp 1–6
83. Velastin SA, Gómez-Lira DA (2017) People detection and pose classification inside a moving train using computer vision. In: International visual informatics conference, pp 319–330 Springer
84. Luo W, Liu W, Lian D, Tang J, Duan L, Peng X, Gao S (2021) Video anomaly detection with sparse coding inspired deep neural networks. IEEE Trans Pattern Anal Mach Intell 43(3):1070–1084
85. Zendel O, Murschitz M, Humenberger M, Herzner W (2017) How good is my test data? introducing safety analysis for computer vision. Int J Comput Vision 125(1–3):95–109
86. Xu K, Jiang X, Sun T (2018) Anomaly detection based on stacked sparse coding with intraframe classification strategy. IEEE Trans Multimed 20(5):1062–1074
87. Ribnick E, Atev S, Masoud O, Papanikolopoulos N, Voyles R (2006) Real-time detection of camera tampering. In: Proceedings of the IEEE international conference on video and signal based surveillance, pp 10–10
88. Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F (2018) Learning from Imbalanced Data Sets vol 10 Springer, Cham
89. Ng A (2021) A Chat with Andrew on MLOps: From Model-centric to Data-centric AI. DeepLearningAI. https://www.youtube.com/watch?v=06-AZXmwHjo
90. Zhao Y, Deng B, Shen C, Liu Y, Lu H, Hua X-S (2017) Spatio-temporal autoencoder for video anomaly detection. In: Proceedings of the 25th ACM international conference on multimedia, pp 1933–1941
91. De Oliveira MF, Levkowitz H (2003) From visual data exploration to visual data mining: A survey. IEEE transactions on visualization and computer graphics 9(3):378–394
92. Tukey JW et al (1977) Exploratory Data Analysis vol 2. Reading, Mass., MA
93. Yazdani M (2016) Using Exploratory Data Analysis to Discover Patterns in Image and Document Collections. PyData Chicago 2016, Pyvideo. https://pyvideo.org/pydata-chicago-2016/using-exploratory-data-analysis-to-discover-patterns-in-image-and-document-collections.html
94. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(6):417–441
95. Jolliffe IT, Cadima J (2016) Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374(2065):20150202
96. Camacho J, Rodríguez-Gómez RA, Saccenti E (2017) Group-wise principal component analysis for exploratory data analysis. J Comput Graph Stat 26(3):501–512
97. Cox MA, Cox TF (2008) Multidimensional scaling. Handbook of data visualization. Springer, Cham, pp 315–347
98. Maaten Lvd, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(Nov): 2579–2605
99. Cox V (2017) Exploratory data analysis. Translating statistics to make decisions. Springer, Cham, pp 47–74
100. Nogami R, Shizuki B, Hosobe H, Tanaka J (2012) An exploratory analysis tool for a long-term video from a stationary camera. In: Proceedings of the IEEE 24th international conference on tools with artificial intelligence, vol 2, pp 32–37
101. Meghdadi AH, Irani P (2013) Interactive exploration of surveillance video through action shot summarization and trajectory visualization. IEEE Trans Vis Comput Grap 19(12):2119–2128
102. Feng Z, Wang J, Harkes J, Pillai P, Satyanarayanan M (2018) EVA: An efficient system for exploratory video analysis. SysML, Indio, California, pp 1–3
103. Matejka J, Glueck M, Bradner E, Hashemi A, Grossman T, Fitzmaurice G (2018) Dream lens: Exploration and visualization of large-scale generative design datasets. In: Proceedings of the 2018 CHI conference on human factors in computing systems, pp 1–12
104. Refaat M (2010) Data Preparation for Data Mining Using SAS. Elsevier, Amsterdam, The Netherlands
105. Stieglitz S, Mirbabaie M, Ross B, Neuberger C (2018) Social media analytics-challenges in topic discovery, data collection, and data preparation. Int J Inf Manag 39:156–168
106. Abati D, Porrello A, Calderara S, Cucchiara R (2019) Latent space autoregression for novelty detection. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 481–490. https://doi.org/10.1109/CVPR.2019.00057

107. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, Van Den Hengel A (2019) Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 1705–1714. https://doi.org/10.1109/ICCV.2019.00179
108. Ionescu RT, Khan FS, Georgescu M-I, Shao L (2019) Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7842–7851
109. Markovitz A, Sharir G, Friedman I, Zelnik-Manor L, Avidan S (2020) Graph embedded pose clustering for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10539–10547
110. Morais R, Le V, Tran T, Saha B, Mansour M, Venkatesh S (2019) Learning regularity in skeleton trajectories for anomaly detection in videos. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11996–12004
111. 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
112. Meher CK, Nayak R, Pati UC (2022) Video anomaly detection using variational autoencoder. In: 2022 IEEE 2nd international symposium on sustainable energy, signal processing and cyber security (iSSSC), pp 1–6. https://doi.org/10.1109/iSSSC56467.2022.10051511
113. Park H, Noh J, Ham B (2020) Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14372–14381
114. Sabokrou M, Khalooei M, Fathy M, Adeli E (2018) Adversarially learned one-class classifier for novelty detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3379–3388
115. Meher CK, Nayak R, Pati UC (2022) Dual stream variational autoencoder for video anomaly detection in single scene videos. In: 2022 2nd Odisha international conference on electrical power engineering, communication and computing technology (ODICON), pp 1–6. https://doi.org/10.1109/ODICON54453.2022.10010086
116. Wang T, Qiao M, Lin Z, Li C, Snoussi H, Liu Z, Choi C (2018) Generative neural networks for anomaly detection in crowded scenes. IEEE Trans Inf Foren Sec 14(5):1390–1399
117. 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
118. Ganokratanaa T, Aramvith S, Sebe N (2022) Video anomaly detection using deep residual-spatiotemporal translation network. Pattern Recogn Lett 155:143–150
119. Zhong J-X, Li N, Kong W, Liu S, Li TH, Li G (2019) Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1237–1246
120. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25:1097–1105
121. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition.arXiv:1409.1556
122. 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
123. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
124. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017)Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
125. Huang G, Liu S, Maaten L, Weinberger KQ (2018) CondenseNet: An efficient densenet using learned group convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition,pp 2752–2761
126. Griffin BA, Corso JJ (2019) Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8914–8923
127. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning.J Big Data 6(1):1–48
128. Narasimhan MG, Kamath S (2018) Dynamic video anomaly detection and localization using sparse denoising autoencoders. Multimed Tools Appl 77(11):13173–13195
129. Yuan H, Cai Z, Zhou H, Wang Y, Chen X (2021) Transanomaly: Video anomaly detection using video vision transformer. IEEE Access 9:123977–123986

130. Nguyen T-N, Roy S, Meunier J (2022) Smithnet: Strictness on motion-texture coherence for anomaly detection. IEEE Trans Neural Netw Learn Syst 33(6):2287–2300. https://doi.org/10.1109/TNNLS.2021.3116212
131. Zhang W, Wang G, Huang M, Wang H, Wen S (2021) Generative adversarial networks for abnormal event detection in videos based on self-attention mechanism. IEEE Access 9:124847–124860
132. Mansour RF, Escorcia-Gutierrez J, Gamarra M, Villanueva JA, Leal N (2021) Intelligent video anomaly detection and classification using faster rcnn with deep reinforcement learning model. Image Vision Comput 104229(2021)
133. Zhang Q, Feng G, Wu H (2022) Surveillance video anomaly detection via non-local u-net frame prediction. Multimed Tools Appl, 1–16
134. Zhang Y, Nie X, He R, Chen M, Yin Y (2021) Normality learning in multispace for video anomaly detection. IEEE Trans Circ Syst Video Technol 31(9):3694–3706. https://doi.org/10.1109/TCSVT.2020.3039798
135. Kukačka J, Golkov V, Cremers D (2017) Regularization for deep learning: A taxonomy. arXiv:1710.10686
136. Jabbar H, Khan RZ (2015) Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices, pp 163–172
137. Lawrence S, Giles CL, Tsoi AC (1997) Lessons in neural network training: Overfitting may be harder than expected. In: AAAI/IAAI, pp 540–545 Citeseer
138. Ruppert D, Carroll RJ (2000) Theory & methods: Spatially-adaptive penalties for spline fitting. Australian & New Zealand J Stat 42(2):205–223
139. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
140. Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C (2015) Efficient object localization using convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 648–656
141. Ko K-E, Sim K-B (2018) Deep convolutional framework for abnormal behavior detection in a smart surveillance system. Eng Appl Artif Intell 67, 226–234
142. Li T, Chen X, Zhu F, Zhang Z, Yan H (2021) Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection. Neurocomput 439:256–270
143. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the international conference on machine learning, pp 448–456 PMLR
144. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
145. Shao L, Zhu F, Li X (2014) Transfer learning for visual categorization: A survey. IEEE Trans Neural Netwo Learn Syst 26(5):1019–1034
146. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Proceedings of the 27th international conference on neural information processing systems - Volume 2. NIPS'14, pp 3320–3328. MIT Press, Cambridge, MA, USA
147. Asad M, Yang J, Tu E, Chen L, He X (2021) Anomaly3d: Video anomaly detection based on 3d-normality clusters. J Visual Comm Image Rep 75:103047
148. Li N, Chang F, Liu C (2021) Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. IEEE Trans Multimed 23:203–215. https://doi.org/10.1109/TMM.2020.2984093
149. Erhan D, Courville A, Bengio Y, Vincent P (2010) Why does unsupervised pre-training help deep learning? In: Proceedings of the Thirteenth international conference on artificial intelligence and statistics, pp 201–208 JMLR Workshop and Conference Proceedings
150. Xian Y, Lampert CH, Schiele B, Akata Z (2018) Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans Pattern Anal Mach Intell 41(9):2251–2265
151. Palatucci M, Pomerleau D, Hinton GE, Mitchell TM (2009) Zero-shot learning with semantic output codes. In: Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A (eds) Advances in neural information processing systems, vol 22. Curran Associates Inc, Canada
152. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1701–1708
153. Koch G, Zemel R, Salakhutdinov R et al (2015) Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop, vol 2 Lille
154. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
155. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp 1532–1543

156. Ullah A, Muhammad K, Haydarov K, Haq IU, Lee M, Baik SW (2020) One-shot learning for surveillance anomaly recognition using siamese 3d cnn. In: 2020 international joint conference on neural networks (IJCNN), pp 1–8 IEEE
157. Xu X, Gong S, Hospedales TM (2017) Chapter 15 - zero-shot crowd behavior recognition. In: Murino V, Cristani M, Shah S, Savarese S (eds) Group and crowd behavior for computer vision, pp 341–369. Academic Press, Cambridge, Massachusetts, United States. https://doi.org/10.1016/B978-0-12-809276-7.00018-7
158. Ramírez Rivera A, Khan A, Bekkouch IEI, Sheikh TS (2022) Anomaly detection based on zero-shot outlier synthesis and hierarchical feature distillation. IEEE Trans Neural Netw Learn Syst 33(1):281–291. https://doi.org/10.1109/TNNLS.2020.3027667
159. Zhou JT, Fang M, Zhang H, Gong C, Peng X, Cao Z, Goh RSM (2019) Learning with annotation of various degrees. IEEE Trans Neural Netw Learn Syst 30(9):2794–2804
160. DeVries T, Taylor GW (2017) Dataset augmentation in feature space. arXiv:1702.05538
161. Gatys LA, Ecker AS, Bethge M (2015) A neural algorithm of artistic style. arXiv:1508.06576
162. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, pp 694–711 Springer
163. Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2574–2582
164. Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, Hammers A, Dickie DA, Hernández MV, Wardlaw J, Rueckert D (2018) Gan augmentation: Augmenting training data using generative adversarial networks. arXiv:1810.10863
165. Wang J, Perez L et al (2017) The effectiveness of data augmentation in image classification using deep learning. Conv Neural Netw Vis Recogn 11:1–8
166. Lemley J, Bazrafkan S, Corcoran P (2017) Smart augmentation learning an optimal data augmentation strategy. IEEE Access 5:5858–5869
167. Cubuk ED, Zoph B, Mane D, Vasudevan V, Le QV (2019) Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 113–123
168. Sutton RS, Barto AG (2018) Reinforcement Learning: An Introduction. MIT press, Cambridge, MA
169. Hernandez-Lopez FJ, Rivera M (2015) AVScreen: a real-time video augmentation method. J Real-Time Image Proc 10(2):453–465
170. Seo Y, Ahn M-H, Hong KS (1998) Video augmentation by image-based rendering under the perspective camera model. In: Proceedings of the Fourteenth IEEE international conference on pattern recognition (Cat. No. 98EX170), vol 2, pp 1694–1696
171. Badrinarayanan V, Budvytis I, Cipolla R (2013) Semi-supervised video segmentation using tree structured graphical models. IEEE Trans Pattern Anal Mach Intell 35(11):2751–2764
172. Budvytis I, Sauer P, Roddick T, Breen K, Cipolla R (2017) Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In: Proceedings of the IEEE international conference on computer vision workshops, pp 230–237
173. Zhang Y, Jia G, Chen L, Zhang M, Yong J (2019) Self-paced video data augmentation with dynamic images generated by generative adversarial networks. arXiv:1909.12929
174. Yun S, Oh SJ, Heo B, Han D, Kim J (2020) Videomix: Rethinking data augmentation for video classification. arXiv:2012.03457
175. Lim SK, Loo Y, Tran N-T, Cheung N-M, Roig G, Elovici Y (2018) Doping: Generative data augmentation for unsupervised anomaly detection with gan. In: Proceedings of the IEEE international conference on data mining (ICDM), pp 1122–1127
176. Joshi A, Namboodiri VP (2019) Unsupervised synthesis of anomalies in videos: Transforming the normal. In: Proceedings of the IEEE international joint conference on neural networks (IJCNN), pp 1–8
177. Radosavovic I, Dollár P, Girshick R, Gkioxari G, He K (2018) Data distillation: Towards omni-supervised learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4119–4128
178. Apat HK, Nayak R, Sahoo B (2023) A comprehensive review on internet of things application placement in fog computing environment. Internet of Things 23:100866. https://doi.org/10.1016/j.iot.2023.100866
179. Wei D, Liu Y, Zhu X, Liu J, Zeng X (2022) Msaf: Multimodal supervise-attention enhanced fusion for video anomaly detection. IEEE Signal Proc Lett 29:2178–2182. https://doi.org/10.1109/LSP.2022.3216500
180. Slavic G, Alemaw AS, Marcenaro L, Regazzoni C (2021) Learning of linear video prediction models in a multi-modal framework for anomaly detection. In: 2021 IEEE international conference on image processing (ICIP), pp 1569–1573. https://doi.org/10.1109/ICIP42928.2021.9506049

181. Kwak YH, Lee J (2021) Toward sustainable smart city: Lessons from 20 years of korean programs. IEEE Trans Eng Manag 1–15. https://doi.org/10.1109/TEM.2021.3060956
182. Yahaya SW, Lotfi A, Mahmud M (2021) Towards a data-driven adaptive anomaly detection system for human activity. Pattern Recogn Lett 145:200–207
183. Santhosh KK, Dogra DP, Roy PP (2020) Anomaly detection in road traffic using visual surveillance: A survey. ACM Comput Surv (CSUR) 53(6):1–26
184. Zhu L, Yu FR, Wang Y, Ning B, Tang T (2018) Big data analytics in intelligent transportation systems: A survey. IEEE Trans Intell Trans Syst 20(1):383–398
185. El-Wakeel AS, Osman A, Zorba N, Hassanein HS, Noureldin A (2019) Robust positioning for road information services in challenging environments. IEEE Sensors Journal 20(6):3182–3195
186. Ren J, Xia F, Liu Y, Lee I (2021) Deep video anomaly detection: Opportunities and challenges. arXiv:2110.05086
187. Stapleton P, Blanchard J (2021) Remote proctoring: Expanding reliability and trust. In: Proceedings of the 52nd ACM technical symposium on computer science education, pp 1243–1243
188. Bawarith R, Basuhail A, Fattouh A, Gamalel-Din S (2017) E-exam cheating detection system.Int J Adv Comput Sci Appl 8(4):176–181
189. Tiong LCO, Lee HJ (2021) E-cheating prevention measures: Detection of cheating at online examinations using deep learning approach–a case study. arXiv:2101.09841
190. Nayak R, Behera MM, Pati UC, Das SK (2019) Video-based real-time intrusion detection system using deep-learning for smart city applications. In: Proceedings of the IEEE international conference on advanced networks and telecommunications systems (IEEE ANTS), pp 1–6 IEEE
191. Nayak R, Behera MM, Girish V, Pati UC, Das SK (2019) Deep learning based loitering detection system using multi-camera video surveillance network. In: Proceedings of the IEEE international symposium on smart electronic systems (iSES)(Formerly iNiS), pp 215–220 IEEE
192. Pandya S, Ghayvat H, Kotecha K, Awais M, Akbarzadeh S, Gope P, Mukhopadhyay SC, Chen W (2018) Smart home anti-theft system: a novel approach for near real-time monitoring and smart home security for wellness protocol. Appl Syst Innov 1(4):42
193. Jayashri S et al (2021) Video analytics on social distancing and detecting mask. Turkish J Comput Math Educ (TURCOMAT) 12(9):2916–2921
194. Hou YC, Baharuddin MZ, Yussof S, Dzulkifly S (2021) Social distancing detection with deep learning model. In: 2020 8th international conference on information technology and multimedia (ICIMU), pp 334–338 IEEE
195. Bhambani K, Jain T, Sultanpure KA (2020) Real-time face mask and social distancing violation detection system using yolo. In: 2020 IEEE bangalore humanitarian technology conference (B-HTC), pp 1–6 IEEE
196. Zuo F, Gao J, Kurkcu A, Yang H, Ozbay K, Ma Q (2021) Reference-free video-to-real distance approximation-based urban social distancing analytics amid covid-19 pandemic. J Transp Health 21:101032
197. Saponara S, Elhanashi A, Gagliardi A (2021) Implementing a real-time, ai-based, people detection and social distancing measuring system for covid-19. J Real-Time Image Proc, 1–11
198. Yu S, Xia F, Sun Y, Tang T, Yan X, Lee I (2020) Detecting outlier patterns with query-based artificially generated searching conditions. IEEE Trans Comput Soc Syst 8(1):134–147
199. Huang H, Yang L, Wang Y, Xu X, Lu Y (2021) Digital twin-driven online anomaly detection for an automation system based on edge intelligence. J Manu Syst 59:138–150
200. He Y, Guo J, Zheng X (2018) From surveillance to digital twin: Challenges and recent advances of signal processing for industrial internet of things. IEEE Signal Proc Mag 35(5):120–129
201. Castellani A, Schmitt S, Squartini S (2020) Real-world anomaly detection by using digital twin systems and weakly supervised learning. IEEE Trans Ind Inf 17(7):4733–4742
202. Fan Y, Wen G, Li D, Qiu S, Levine MD, Xiao F (2020) Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder. Comp Vision Image Underst 195:102920
203. Li B, Leroux S, Simoens P (2021) Decoupled appearance and motion learning for efficient anomaly detection in surveillance video. Comp Vision Image Underst 210:103249
204. Wu J-C, Lu S, Fuh C-S, Liu T-L (2021) One-class anomaly detection via novelty normalization. Comp Vision Image Underst 210:103226. https://doi.org/10.1016/j.cviu.2021.103226
205. Wang J, Xu Z (2016) Spatio-temporal texture modelling for real-time crowd anomaly detection. Comp Vision Image Underst 144:177–187
206. Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp 41–48
207. Mao X, Li Q, Xie H, Lau RY, Wang Z, Paul Smolley S (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2794–2802

208. Hacohen G, Weinshall D (2019) On the power of curriculum learning in training deep networks. In: Proceedings of the international conference on machine learning, pp 2535–2544 PMLR
209. Shen J, Tao D, Li X (2008) Modality mixture projections for semantic video event detection. IEEE Trans Circ Syst Video Technol 18(11):1587–1596. https://doi.org/10.1109/TCSVT.2008.2005607
210. Wu R, Yan S, Shan Y, Dang Q, Sun G (2015) Deep image: Scaling up image recognition 7(8)(2015) . arXiv:1501.02876
211. Zhao A, Balakrishnan G, Durand F, Guttag JV, Dalca AV (2019) Data augmentation using learned transformations for one-shot medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8543–8553
212. Shi Y, Wei Z, Ling H, Wang Z, Shen J, Li P (2021) Person retrieval in surveillance videos via deep attribute mining and reasoning. IEEE Trans Multimed 23:4376–4387. https://doi.org/10.1109/TMM.2020.3042068
213. Chilimbi T, Suzue Y, Apacible J, Kalyanaraman K (2014) Project adam: Building an efficient and scalable deep learning training system. In: 11th $\{USENIX\}$ symposium on operating systems design and implementation ($OSDI$ 14), pp 571–582
214. Liu M, Zhao J, Zhou Y, Zhu H, Yao R, Chen Y (2022) Survey for person re-identification based on coarse-to-fine feature learning. Multimed Tools Appl 81(15):21939–21973

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.