Multimedia Tools and Applications (2024) 83:27761–27807
https://doi.org/10.1007/s11042-023-16445-z

A critical study on the recent deep learning based semi-supervised video anomaly detection methods

Mohammad Baradaran[1] · **Robert Bergevin**[1]

Received: 29 November 2022 / Revised: 3 July 2023 / Accepted: 2 August 2023 /
Published online: 19 August 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Video anomaly detection (VAD) is currently a trending research area within computer vision, given that anomalies form a key detection objective in surveillance systems, often requiring immediate responses. The primary challenges associated with video anomaly detection tasks stem from the scarcity of anomaly samples and the context-dependent nature of anomaly definitions. In light of the limited availability of labeled data for training (specifically, a shortage of labeled data for abnormalities), there has been a growing interest in semi-supervised anomaly detection methods. These techniques work by identifying anomalies through the detection of deviations from normal patterns. This paper provides a new perspective to researchers in the field, by categorizing semi-supervised VAD methods according to the proxy task type they employ to model normal data and consequently to detect anomalies. It also reviews recent deep learning based semi-supervised VAD methods, emphasizing their common tactic of slightly overfitting their models on normal data using a proxy task to detect anomalies. Our goal is to help researchers develop more effective video anomaly detection methods. As the selection of a right Deep Neural Network (DNN) plays an important role in several parts of this task, a quick comparative review on DNNs is also included. Unlike previous surveys, DNNs are reviewed from a spatiotemporal feature extraction viewpoint, customized for video anomaly detection. This part of the review can help researchers select suitable networks for different parts of their methods. The review provides a novel and deep look at existing methods and results in stating the shortcomings of these approaches, which can be a hint for future works.

Keywords Video anomaly detection · Spatio-temporal feature extraction · Deep learning · Semi-supervised · Proxy-task

✉ Mohammad Baradaran
mohammad.baradaran.1@ulaval.ca

Robert Bergevin
robert.bergevin@gel.ulaval.ca

[1] Department of Electrical and Computer Engineering, Université Laval, 1065 Av. de la Médecine, Québec G1V 0A6, QC, Canada

1 Introduction

Anomaly Detection (AD) is one of the essential and crucial tasks in various applications, such as video surveillance, quality control in production lines, security systems in data transmissions, etc. Anomaly detection (a.k.a., abnormal event detection or outlier detection) involves detecting patterns in data (image, video, etc.) that do not conform to expected behavior or the notion of normal behavior (i.e., behavior conformed by the majority of data samples) [19]. In video anomaly detection, the goal is to precisely locate the anomalies (spatially and temporally) inside frame sequences. Anomalies may be of different types, but they generally share these assumptions: 1- Anomalies rarely take place (compared to normal events), so they have a low probability of occurrence. 2- Patterns of anomalies are distinct from normals. These assumptions are the keys to identifying anomalies, however, detecting anomalies is generally challenging for a number of reasons:

1- There is not a limited and precise definition for abnormality. Anomaly patterns are diverse and unrestricted and hence cannot be modeled or predicted precisely.
2- The boundary between normals and anomalies is not often precisely defined. Besides, it is hard to classify the data instances near this boundary.
3- Abnormalities are highly contextual and their definitions can change considering the time, place, and environment. For example, driving a car at a speed of 100 km/h is a normal behavior on a highway but it cannot be considered as normal, in a residential area.
4- Anomalies are rare (but diverse) and there are not enough labeled anomaly samples to train a model.
5- It is very difficult to define a precise boundary (model) around normals, which can cover all normal patterns and behaviors.
6- The most complex challenge would be intelligent anomalies (adversarial samples) that attempt to resemble normal patterns.

1.1 Video Anomaly Detection (VAD)

Video anomaly detection has the same definition as mentioned above, but here we deal with videos and we strive to detect anomalous video events, spatially and temporally. Hence, in video, appearance and motion features are the key elements for defining anomalies and they should be extracted effectively and analyzed jointly. For video anomaly detection, we have the same general anomaly detection challenges (as mentioned above). There are some additional challenges, related to video analysis, such as high dimensionality of video data, complex scenes, occlusions, high interaction inside video contents, low resolution, etc.

1.2 Categories of video anomalies

Depending on the problem (definition and context), anomalous data can fall into one of these three sub-categories: point anomalies, contextual (conditional) anomalies, or collective (group) anomalies [19].

For point anomalies, the anomaly is defined and recognized, by analyzing the value of one single data instance, individually. This value can be a scalar or a feature vector. For example, in a video, a frame can be labeled as an anomaly, simply by detecting an unexpected object (appearance feature vector) or by capturing a vehicle, which is moving with a speed greater

than the allowed speed (regardless of the vehicle type or the place). In contrast, for conditional anomalies, contextual information is required, in addition to the value. In this case, a single factor (a variable value for instance) is not enough to make a robust and careful decision. For example, it is expected to see cars in the street, but in their designated lines, not on the sidewalk. As another example, although 100 km/h is an allowed speed on highways, it is considered to be an abnormal behavior on snowy slippy roads. As noted, in these examples, the values of the features are not individually enough and they can be interpreted differently, in different conditions (depending on the context). Finally, in collective anomalies, groups of data instances form the anomalies. For example, the presence of one or a few people may be normal in the bank, but a group of one hundred people would be considered an anomaly.

Video anomalies can fall into any of the discussed categories, though the specific classification generally depends on factors such as the problem's definition, complexity, context, or the objective of simplifying the problem. This means that formulating a detection method based on any of these categories can yield certain benefits or limitations. For instance, in the case of object-centric VAD methods (discussed in Section 3.6), extracting and analyzing objects independently from their frames enhances focus on the object, but may lead to the omission of some contextual information, such as the object's location within the scene.

1.3 Effective video anomaly detection

Regarding the definition, type, and challenges of the video anomaly, the following items should be considered, to have an effective computer vision system for video anomaly detection.

1: Precisely defining normality (normal patterns).
2: Extracting effective and discriminative spatiotemporal features, customized for the given task. This point is thoroughly examined in Section 2.
3: Considering the differences and similarities in and between normal and abnormal behaviors.
4: Considering environment information (e.g., illumination changes) and its variations. Our experiments in Section 4 highlight this point.

Anomaly detection methods are generally divided into supervised, semi-supervised, unsupervised, and weakly supervised. Semi-supervised methods have gained more attention since there is not enough labeled data for anomalies (as they are rare and diverse, and it is hard to collect enough samples to cover all anomalies). On the other hand, not only is there enough training data for normals for semi-supervised formulation but also the definition of anomalies is tied with the definition of normality (which is the base of semi-supervised VAD). This paper primarily concentrates on providing a comprehensive review of semi-supervised VAD approaches.

Important factors concerning semi-supervised anomaly detection that should be considered are:

1) Availability of enough labeled data for normalities to cover all of the normal patterns [134].
2) Extracting compact and discriminative features for normal patterns, to ensure that normal features are very similar and close to each other and very distinct from features of anomalies.

In this study, we have endeavored to take into account these considerations both in our review of various VAD methods and in our experimental work.

1.4 Benchmark datasets

UCSD (ped1 and ped2) [75] is one of the most popular datasets in semi-supervised video anomaly detection, in which the normal scenes include people walking in the walkways, while anomalies are due to the presence of unexpected objects in the scene (such as carts, bicycles, skateboards, etc.), different motion patterns (skateboard riding, etc.) and walking in the grass. The main challenge of this dataset is the low resolution of the frames. The Shanghai Tech dataset [64] is a similar dataset to UCSD, considering the definition of normals and anomalies. However, unlike The UCSD dataset, its resolution is high and it would be easier for systems to recognize anomalous objects, using appearance-based features. The video clips of this dataset are captured across diverse campus scenarios. In these scenarios, the normal instances depict individuals engaged in regular activities like walking. Conversely, the anomalies in the dataset arise from the presence of unfamiliar objects or abnormal activities, such as running, chasing, jumping, and so on.

The CUHK Avenue dataset [73], UMN dataset [2] and Subway dataset [5] consider the activities of people in crowded scenes to define normality. The Street Scene dataset [51] is a recently proposed high-resolution dataset. Within this dataset, the definition of normality varies depending on the objects involved, such as humans, cars, bikes, and others. To illustrate, walking along a sidewalk is considered a normal activity for people. However, crossing the road or engaging in unusual activities on the sidewalk, such as loitering, would be classified as anomalies in this context. Among all mentioned datasets, the Street Scene dataset is the most challenging, since 1) the events, inside, are highly contextual (in addition to the appearance and motion information, the location information of the objects is essential in defining the anomalies). Figure 1 illustrates this point. 2) The anomalies are of various types and they are numerous. The mentioned datasets are compared in detail, in Tables 1 and 2.

It is important to note that the UCF-Crime dataset [114], originally designed and utilized for activity recognition tasks and supervised video anomaly detection methods, can also be employed to evaluate semi-supervised video anomaly detection approaches. This dataset comprises videos depicting a range of criminal activities, including theft, burglary, and robbery, along with a category for normal activities. In this context, the "Normal" category serves as the representation of typical patterns, while other categories such as shooting or robbery are considered anomalies.

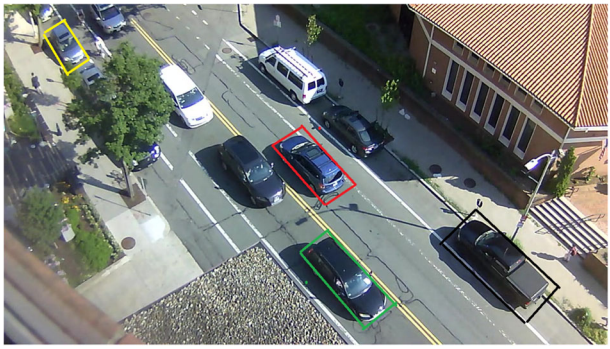

Fig. 1 A frame from the Street Scene dataset [51]. Anomalies in this dataset are highly contextual. For example, the definition of anomaly is different, in these 4 positions (4 different boxes), for the same class of object. Best viewed in color

Table 1 Detailed information on benchmark video anomaly detection datasets

Datasets	Training frames	Test frames	Resolution	NO. of anomalies	Annotation	Color	No. of scenes	Format
UCSD-ped1	6,800	7,200	238 x 158	54 (5 types)	Pixel level-binary mask	gray	1	tif
UCSD-ped2	2,550	2,010	360 x 240	23 (5 types)	Pixel level- binary mask	gray	1	tif
ShanghaiTech	274,515	42,883	856 x 480	130	Pixel level- binary mask	color	13	avi
Street scene	56,847	146,410	1280 x 720	205 (17 types)	Pixel level- bounding box	color	1	jpg
CUHK Avenue	15,328	15,324	640 x 360	47 (5 types)	Pixel level-mask	color	1	tif
UMN	7,740 total frames	–	240 x 320	11 (1 type)	Frame level	color/ gray	3	avi
Subway (Entrance)	18,000	68,535	512 x 384	66 (5 types)	Frame level	gray	1	avi,tif
Subway (Exit)	4,500	34,440	512 x 384	19 (3 types)	Frame level	gray	1	avi,tif
UCF-Crime	1,190 clips x 7247 frames	810 clips x 7247 frames	320 x 240	13 types	Frame level	color	multiple	mp4

Table 2 Comparing existing benchmark VAD datasets, based on their definition of anomaly and their challenges

Datasets	Anomalies (some)	challenges/ special points
UCSD	Non-pedestrian entities in walkways (bikes, skates, small carts, wheelchair), unfamiliar motion patterns (people walking across the walkway or on the grass).	*The definition of anomaly is similar in Ped1 and Ped2. *Different scales for different distances. *Object types are not always recognizable (due to resolution and distance).
ShanghaiTech Campus	Sudden motion, such as chasing and brawling, unexpected objects.	*Multiple scenes with multiple view angles. *Complex lighting conditions.
Street scene	Jaywalking, loitering, a car outside lane, car u-turn, car illegally parked, biker on the sidewalk, etc. (17 types of anomalies)	*The anomalies are highly contextual and more challenging than other datasets *High resolution *The high number of anomaly types. *Presence of minor camera motion in some frames.
CUHK Avenue	Throwing objects, loitering, running.	*The size of people may change because of the camera position and angle *Camera shakes in some frames.
UMN	Crowd escaping quickly from the scene.	*Number of anomalies is limited (just one anomaly type) *The video is short. *Low resolution.
Subway	Moving in the wrong direction, entering without payment, loitering.	*Noisy video *There is a big timer on the screen. *Objects at distance are not clear.
UCF-Crime	Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism	*Real-World Complexity *Diversity of Criminal Activities *Camera Viewpoint Variation

Different items are separated by asterisks (*) in the table

State-of-the-art semi-supervised VAD methods have selected from among these datasets to assess their proposed methods and compare them with other studies. It's important to note that the choice of the dataset for evaluating a VAD method and demonstrating the explainability of its detections should align with the method's formulation and objectives. In essence, the context of the anomaly within a dataset should match the formulation and goal of the VAD method. Studies that use inappropriate datasets may fail to contribute any valuable insights.

1.5 Other surveys

There are a few other survey articles published concerning anomaly detection. For example, Ramachandra et al. [91] review single-view video anomaly detection methods, with special consideration of the applicability of methods on currently available benchmark datasets. They discuss the problem formulation in distance-based, probabilistic, and reconstruction-based VAD methods and compare the performance of state-of-the-art methods quantitatively. Kiran et al. [55] review the state-of-the-art Deep Learning (DL) based approaches for anomaly detection in videos and categorize them based on the criteria of detection and the type of network used. It reviews supervised and semi-supervised VAD methods such as reconstruction-based and prediction-based models for anomaly detection and hence is similar to our work, in some parts. However, the deep networks are studied mostly focusing

on their structures and basic concepts, but not from a feature extraction viewpoint and their compatibility with the anomaly detection task. Moreover, Our work reviews and analyzes numerous proxy tasks used in the filed. Raghavendra et al. [90] present a general review on DL-based anomaly detection methods and review their applicability in different fields of application (i.e. Cyber-Intrusion Detection, Medical Anomaly Detection, Sensor Networks Anomaly Detection, Internet Of Things (IoT) Big-data Anomaly Detection, Log-Anomaly Detection, Video Surveillance, Industrial Damage Detection). This review is an application-based categorization and mostly summarizes the Deep Neural Network (DNN) types, used for various applications. Chandola et al. [19] review Machine Learning based anomaly detection methods (conventional Machine Learning (ML) approaches) based on the different pattern recognition techniques used (such as clustering, classification, neural networks, etc.) and it studies them for different applications. Moreover, this review explains the basic assumption, advantages, computational cost, etc, for each of the techniques. Bulusu [17] has provided a review on DL-based anomalous instance detection methods. Its focus is on discussing unintentional and intentional anomalies, specifically in the context of DNNs. Shibin [108] reviews evaluation metrics and popular evaluation schemes, used to measure the performance of video and image anomaly detection approaches. Table 3 summarizes existing surveys, based on the subjects they have covered.

Furthermore, a number of short reviews have been published within the field. Ren et al. [95] provide an overview of the possibilities (including use cases in public health) and challenges associated with deep learning-based VAD models, such as reconstruction-based, prediction-based, generative, and hybrid methods. Yadav et al. [129] summarize recent methods, primarily focusing on weakly-supervised approaches, as well as the evaluation metrics employed. They also discuss the current trends in the field. Lastly, Roka et al. [98] delve into a comprehensive analysis of the advantages and disadvantages of various machine-learning and non-machine-learning techniques, with a particular emphasis on the application of GANs in VAD.

In this study, our focus lies on semi-supervised Video Anomaly Detection (VAD) methods, and we approach these works from a unique perspective - specifically, we examine how they employ different proxy tasks to effectively model normal patterns. Unlike other studies, our analysis assesses their performance based on their success in analyzing both motion and appearance patterns. Additionally, we endeavor to highlight the strengths and weaknesses of different VAD methods. Finally, diverging from other reviews that simply list different deep neural networks used in VAD methods, we review them from a fresh angle: effective spatio-temporal feature extraction. This perspective tailors the review of DNNs specifically for video anomaly detection and can assist researchers in selecting the appropriate network type for different stages of their method.

1.6 Topic and contributions

In this study:

- Deep Neural Networks are evaluated and contrasted with a focus on spatiotemporal feature extraction and pattern learning. This fresh perspective is advantageous in video anomaly detection. It assists researchers in choosing the most appropriate network for various components of their methodology, taking into account the specific objectives of their anomaly detection method.
- Recent deep learning-based semi-supervised anomaly detection methods are examined and contrasted, with an emphasis on their strengths and weaknesses. This perspective

Table 3 Comparing existing AD and VAD surveys, based on their contents and subjects covered

Reference	[1]	[8]	[9]	[10]	[11]	[12]	Our study
Reconstruction		✓	✓				✓
Prediction			✓				✓
Object centric							✓
Segmentation							✓
Memorization							✓
ST feature extraction			✓				✓
Datasets		✓					✓
Evaluation metrics		✓	✓			✓	✓
DNN/ ML approaches	ML	DNN/ML	DNN	DNN	DNN/ML		DNNs
Focused applications	Intrusion detection, fraud detection, medical anomaly detection, industrial damage detection, image processing, text, sensor networks.	Video anomalies	Video anomalies	Fraud detection, cyber intrusion detection, IoT, video, industrial damage, sensor, etc.	Fraud detection, malware detection, healthcare, video surveillance, etc.	Image and video anomaly detection	Video anomalies
Topics covered	*Anomaly detection in different applications. *Different ML approaches for anomaly detection.	*Distance-based, probabilistic and reconstruction-based anomaly detection approaches.	*Reconstruction models, predictive models, generative models for AD.	Semi-supervised, unsupervised, hybrid models, one class neural networks for AD.	*Detection of unintentional anomalies, *Detection of intentional anomalies*Applications	*Evaluation schemes. *Evaluation metrics.	*DNNs from a feature extraction viewpoint. *AD methods based on their ST feature extraction processes. *Semi-supervised AD methods.

DNN stands for Deep Neural Network, DL for Deep Learning, ST for Spatio-Temporal, and ML for Machine Learning. Different items are separated by asterisks (*) in the table

can provide valuable insights for future research and assist researchers in selecting the most suitable strategy, tailored to their specific objectives.
– Common aspects of all recent DL-based semi-supervised anomaly detection approaches (especially, their implicitly common strategy for anomaly detection) are stated. This provides a new, global, and integrated perspective to the field.
– Selected experiments have been performed to highlight the strengths and weaknesses of certain video anomaly detection methods. These findings shed light on areas requiring further research and can guide future researchers in considering these aspects in their work.

1.7 Organization of the paper

Section 2 provides a comprehensive review of deep neural networks, specifically examining their efficacy in spatio-temporal feature extraction. The analysis concentrates on their appropriateness for different components within the video anomaly detection method, such as modeling motion or appearance. In Section 3, a general look at different anomaly detection approaches is provided, and state-of-the-art DL-based semi-supervised anomaly detection methods are reviewed and compared, based on how they formulate and address the problem. Shortcomings of existing methods are also listed in this section, which can be the subject of future work. Finally, experiments are conducted in Section 4 to clarify the shortcomings of the existing methods. The structure of the content addressed in the survey is depicted in Fig. 2.

2 Deep Neural Networks (DNNs)

Like most computer vision tasks, video anomaly detection is completely reliant on effective feature extraction. Hence, it is very important to have a good understanding of DNNs feature

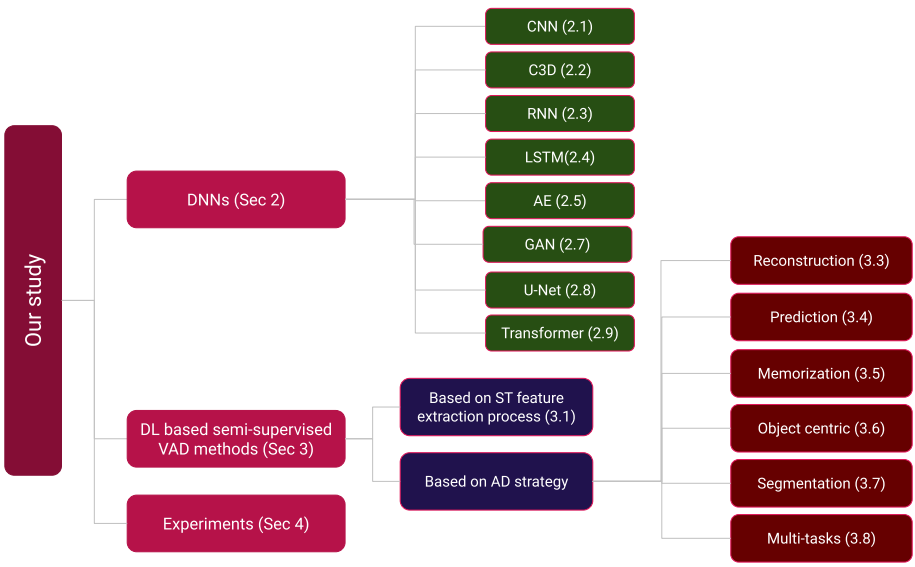

Fig. 2 The content addressed in the study

extraction capabilities, as they are the key tool for feature extraction and pattern learning. In this section, a general, yet effective, look at various deep models and their applicability for different related sub-tasks (e.g., modeling motion and/or appearance patterns) is provided. Their compatibility with different data types is also analyzed, aiming to assist researchers in selecting the most suitable network for their specific needs and data characteristics.

Deep learning has brought great success to various applications and research fields, especially computer vision applications, in analyzing high-dimensional data. DNNs have been useful for different purposes and steps in computer vision applications. More specifically, in video anomaly detection, they have been used to:

1) Extract discriminative high-level Spatio-Temporal features, for different types of data (such as spatial data, sequences, etc.), by using proper architectures (such as CNNs, RNNs, etc.)
2) Learn, model, and memorize patterns and information.
3) Differentiate between normal and abnormal patterns.

Hence, it will be useful to review DNNs, considering the mentioned factors. These networks are analyzed from different points of view such as their architectures, feature extraction ability, compatibility with different data types, and their applicability to different tasks. Additional information about DNNs and their applications can be found in [27].

2.1 Convolutional Neural Networks (CNN)

Convolutional neural networks are special forms of feed-forward neural networks and are composed of multiple convolutional and pooling layers, which are followed by a few Fully Connected (FC) layers, at the end of the network. Unlike fully connected networks, the architecture of CNNs is compatible with 2D structured inputs (such as images or any other 2D signals), which helps effectively preserve the spatial structure of inputs. Feichtenhofer et al. [37], present a deep insight into convolutional neural networks, for video recognition tasks. Convolutional layers are composed of multiple kernels, which are convolved with the input image or mid-layer activation maps to produce next-level activations. Benefiting from some features such as weight sharing mechanism, invariance to translation and local pattern identification makes CNN a good choice for image processing.

In some networks such as Autoencoders, in order to reconstruct an image from extracted features in the latent space, there must be up-sampling-like layers to increase the resolution of feature maps. Transpose convolution, which is also referred to as deconvolution or up-convolution, is a convolution-based operation, which increases the resolution of its input. Up-sampling is a similar operation to transpose-convolution, but the main difference is that transpose-convolution has trainable kernels.

2.1.1 Characteristics of CNNs

Videos are consecutive frames that should be processed both separately (image processing) and also in connection to each other (considering their temporal dependencies). Convolutional neural networks are generally the essential elements for image processing. This is due to some important characteristics of CNNs, which make image processing more effective, efficient and even less challenging. Some of these characteristics are listed as below:

1: Reduced number of parameters: Thanks to local connectivity and shared weights, CNNs have much fewer parameters, compared to Fully Connected Networks, and hence are easier to train.

2: Shift/Translation invariance: this means that by any shift in input, the result does not change (because of convolutional and pooling layers).
3: Transfer Learning: The transfer Learning possibility is one of the strengths of CNNs, in which pre-trained networks are used for feature extraction, in other similar datasets. Nazare et al. [83] study the quality of features, extracted by pre-trained CNNs, for anomaly detection tasks. Ionescu et al. [49] and Aburakhia et al. [4] are good examples of the application of pre-trained CNNs for extracting appearance features to detect anomalies in videos.
4: Convolutional neural networks (CNNs) are suitable for processing input data that has an inherent grid-like topology.
5: CNNs extract rich features at different semantic levels.
6: More filters capture more features but increase the computational cost [24].

2.1.2 CNNs from a feature extraction point of view

As mentioned in the previous subsection, CNNs render image processing (and hence video processing) more efficiently and effectively, due to their ability to exploit spatial features. Therefore they are the prime element for spatial feature extraction from frames. Here are some important aspects to consider regarding CNNs from the feature extraction viewpoint:

1: Experiments show that features in the first layers are low-level and local. For example, filters in the first layer are edge detectors and color filters. The edge detectors are at different angles and allow the network to construct more complex features in the next layers [83].
2: Layers towards the end of the network learn high-level combinations of the features learned in the earlier layers (see Fig. 3).
3: Although the deeper layers have complex and higher-level features and are usually used as feature representation, in order to have a better performance for a special task, it is the target task that precisely defines the layer from which the features should be extracted. For example, in some tasks, such as iris recognition, the recognition accuracy drops after special layers, because the network captures only the abstract and high-level information and it does not distinguish much between diverse iris patterns [8].
4: Reducing the kernel size can improve the capture of smaller details in the picture while missing the global information in the frame and may result in greater confusion. Larger

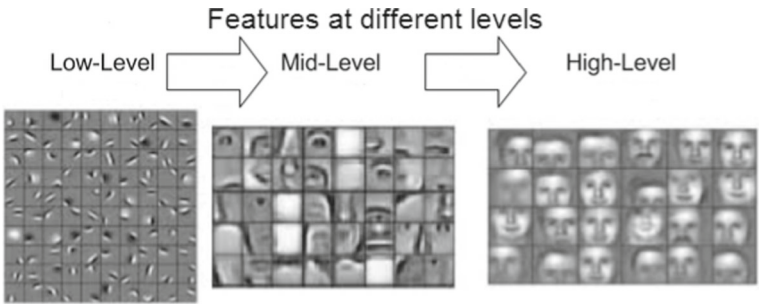

Fig. 3 Different levels of features extracted in CNNs. In this figure, different levels of features are extracted for human face pictures. Left: extracted low-level features are generic and focus on edges. Middle: CNN focuses on different parts of the object at mid-levels. Right: deeper layers provide a global look at objects, extracting high-level features. This figure originally appeared in [60]

kernels, on the other hand, will lead to a global look at the image, while missing the details [53]. This is extremely important when there are objects at different distances from the camera (different scales). Hence, the filter size should be selected considering the task, dataset, and application. Some works, such as [84], utilize inception modules in early layers, to automatically select the proper kernel size.

5: Each kernel is in charge of learning special features from the image. For example, convolutional kernels are capable of capturing features such as edge, line, texture, shape, intensity, color, etc. [105].

6: Earlier layers in a CNN concentrate on generic features (independent of the task), while deeper layers extract features more specific to the problem and the goal.

2.1.3 CNNs for spatiotemporal feature extraction

As explained before, CNNs are excellent and powerful in feature extraction from images. When it comes to consecutive frames (video clips or other 3D tensors), CNNs are not, by nature, suitable for the capture of temporal patterns [77], since they consider single frames as input. Moreover, 2D convolutional kernels map each receptive field (2D or 3D) to one channel (note that, kernel depth in CNNs is equal to the number of input channels). In order to allow the network to be aware of temporal variations, the use of a cuboid of frames (instead of a single frame), as the input, has been proposed. However, the first convolution destroys the temporal structure and does not show promising results in capturing temporal patterns [117].

2.2 3D Convolutional networks (C3D)

In C3D, 3D kernels (with sizes smaller than the width, height, and depth of frame sequences) are applied on consecutive frames and the output is a 3D tensor, unlike the 2D networks which produce a one-channel output for either an image input or sequence of frames (see Fig. 4). The C3D network considers spatial information in the first few frames and it starts to consider the temporal information in the following frames [116].

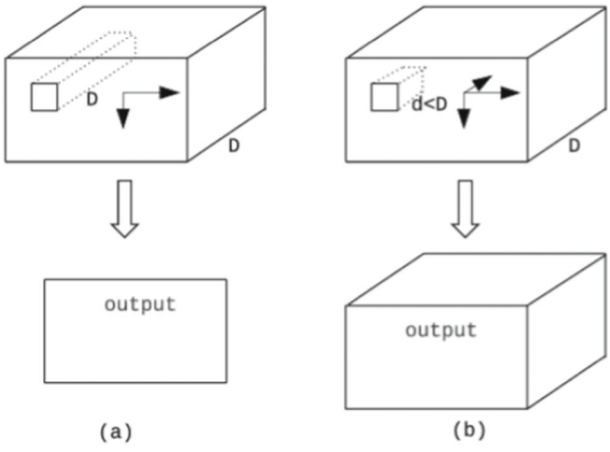

Fig. 4 Different structures in 2D and 3D convolutions and their different output feature maps [119]. (a): in a 2D CNN, the output feature map is a single channel tensor. (b): in a 3D CNN the output is a 3D tensor

2.2.1 Characteristics of C3Ds

The architecture of C3Ds seems to be a good choice for spatiotemporal feature extraction. However, there are some points about C3Ds that should be considered for feature extraction.

1- C3D achieves better results in video analyzing tasks (such as video classification and video retrieval), compared to 2D CNN, as it captures both spatial and temporal information [18].
2: C3D requires a high number of parameters, thus it is computationally expensive and difficult to train, which makes it prone to overfitting [52].
3- Modeling the long sequences is not addressed in C3D, because it leads to a huge computational cost.
4- C3D does not take advantage of Transfer Learning as effectively as 2D CNNs.

2.3 Recurrent Neural Networks (RNNs)

Basic feed-forward networks (such as CNNs) accept a fixed-sized input and produce fixed-sized outputs. This is one of the shortcomings of the feed-forward networks, which are therefore not applicable for some applications, such as language translating or frame captioning, in which the length of the input sentence (or image) and its translation might be different. This problem is addressed by recurrent neural networks. Moreover, unlike a feed-forward network, in which data pass through layers once, in RNNs, they cycle in a loop and touch neurons several times. In this way, RNNs not only consider the current input but also care about its temporal neighbors (past or future frames). More importantly, as Recurrent Neural Networks feature inner loops, they allow the information to persist [112].

2.3.1 Characteristics of RNNs

Recurrent neural networks are, by nature, compatible with sequences [36]. Hence they are widely used for temporal feature extraction. However, regarding the models' abilities and the target task, some points should be considered regarding RNNs, as listed below:

1: As they benefit both new input data and the previous hidden state, as the input to the network, they are able to model sequences and to extract temporal information.
2: Thanks to the presence of the hidden state, they benefit from an internal memory [29].
3: This model is not limited to fixed input and output sizes and hence is appropriate for several tasks such as video captioning, translating, etc.
4: RNNs have difficulties learning long-term dependencies, because of vanishing and exploding gradients [86].

2.4 Long Short-Term Memory Units (LSTMs)

LSTM is a special type of RNN, designed to avoid the long-term dependency problem. LSTMs are gated memory blocks, which include 3 special gates in their chain-like structure, and in addition to hidden states (as is in RNNs), they have cell states. Carefully regulated by gates, LSTM has the ability to remove or add more information to the cell states. Gates, in LSTM, are composed of a sigmoid layer and a pointwise multiplication operation. Since the sigmoid function produces outputs between zero and one, it defines how much information should be deleted or passed [79].

Like every type of neural network, layer size (memory units, here in LSTM) and network depth are the hyper-parameters to choose. Generally, deeper models show better performance in extracting richer features, compared to shallow models, but using much deeper models does not always guarantee the best performance, for all types of applications and tasks.

2.4.1 Characteristics of LSTMs

LSTM has some extra advantages compared to simple RNNs in modeling sequences (for example an inherent memory), which makes it the first choice for sequences, in most cases. However, other practical points, as listed below, should be considered about LSTMs.

1: LSTMs handle exploding and vanishing gradients effectively, thus they are able to model longer sequences, compared to the basic RNN structure, although this length also depends on the nature of the sequence data and its inner correlation [24].
2: Although LSTMs have no difficulty in modeling long dependencies, they lead to high computational complexity, when modeling long sequences.
3: LSTM is the basic element of temporal attention mechanisms.
4: LSTMs have the ability to learn the context required for making predictions in sequence data, therefore they are widely used for forecasting tasks [58].

2.4.2 C3D versus LSTMs, in modeling temporal information

Although both 3D convolutional networks and recurrent neural networks consider sequences and model temporal information, the nature of patterns, captured by these models, are quite different. In LSTMs, based on the task, the network can be encouraged to select meaningful time dependencies and forget unnecessary items. Moreover, the network follows the evolution between sequences. In C3Ds, the network attempts to memorize the patterns inside the training cuboid (frame sequences) without explicitly emphasizing the order of the frames. Moreover, the extracted patterns in C3Ds are more generic [116].

2.4.3 Special points regarding ConvLSTM

LSTMs in their basic form are not suitable for 2D spatially structured data. Hence they are extended to ConvLSTM, in which multiplications are replaced with convolutions.

1: ConvLSTM shows great performance in extraction of spatiotemporal features, by taking advantage of its two main elements: i- LSTM to capture long temporal dependency and ii- Convolution for structured spatial information.
2: Due to convolution, ConvLSTM is capable of capturing local spatial information and suitable for spatiotemporal localization tasks [92].
3: In LSTM, convolutional kernel size of input-to-input connection, defines the resolution of the feature map produced from the input. In addition, the filter size of hidden-to-hidden connections defines the collective information from previous steps. Moreover, larger transitional kernels capture faster motions, while smaller kernels perceive slower ones [77].

2.4.4 GRU versus LSTM

The Gated Recurrent Network (GRU) is another improved version of the standard recurrent neural network to solve the vanishing gradient problem of a standard RNN. It is a gated

memory block similar to LSTM, with a different number of gates (3 gates in LSTM and 2 in GRU). Chung et al. [25] evaluate the performance of these networks on sequence modeling.

1: GRU has a simpler architecture compared to LSTM and its training is faster [124].
2: In theory, LSTMs can learn longer sequences than GRUs and they perform better for longer sequences.
3: In general, the relative performance of each method depends on the data and the application [25].
4: Unlike GRU, which exposes its full content (seen or used content) without any control, the amount of the memory content is controlled by the output gate, in LSTM [25, 78].

LSTM and GRU have gained significant popularity in prediction tasks. Given that future frame/video prediction serves as a common proxy task in video anomaly detection, these networks are regarded as essential tools for such tasks. However, choosing between LSTM and GRU depends on factors such as the complexity of the data and the specific requirements of the task at hand.

2.5 Autoencoders

A deep Autoencoder is an unsupervised learning network architecture (learning from unlabeled training samples) composed of two main sections, encoder, and decoder, which aims to map input data to a latent space, in order to extract deep features and then reconstruct the input using extracted features. In other words, it attempts to learn an approximation to the identity function, so that the output would be similar to the input. Recently, autoencoders are widely used in anomaly detection (especially video anomaly detection). This is because of their ability in unsupervised representation learning. Here are some points regarding Autoencoders (AEs), which should be considered by researchers, in video anomaly detection:

1: They extract effective representations from data, in an unsupervised approach.
2: Autoencoders are effectively used for noise removal [42].
3: Autoencoders are effectively used for dimensionality reduction similar to PCA. The difference is that PCA is restricted to a linear map, while autoencoders can have nonlinear encoders/decoders [9].
4: The basic Autoencoder consists of fully connected layers and is therefore not only computationally expensive but also unsuitable for image processing, as it flattens the image to a vector and destroys the spatial structure (this problem is addressed by convolutional Autoencoders).
5: A baseline Autoencoder is not complex enough to learn complex information (such as image content), and thus generally attempts to memorize and average the data (this problem is addressed, partially, by Variational Autoencoders).
6: The fundamental problem with Autoencoders, for generation tasks, is that their latent space, may not be continuous, or allow easy interpolation [57].

2.5.1 Shortcomings of deep auto-encoders

As mentioned before, AEs are appealing tools for anomaly detection researchers. However, they sometimes do not produce the desired results, mostly due to these facts:

1: They are prone to vanishing gradients.

2: They reproduce a lower-quality version of the input image, without explicitly considering its high-level contents. Pihlgren et al. [88] introduces perceptual loss instead of element-wise loss to alleviate this shortcoming.
3: Autoencoders confront information imbalance in each layer [64]. This challenge is addressed using U-Nets (see Section 2.8).
4: Autoencoders are unsupervised feature extractors and are not aware of the classes of the objects inside the image.
5: Autoencoders suffer from memorization and their reconstructed images are blurry [13]. GANs address this challenge of Autoencoders [13].

2.6 Variational Autoencoders (VAE)

Variational AE (VAE) is a generative variant of classical AEs, which assumes a probability distribution (such as a Gaussian distribution) for the source input data and it attempts to capture the parameters of the distribution, through an encoding-decoding process. In VAE, not only does the network attempt to reconstruct an image but the network is also asked to consider the same distribution, for the generation of new samples, as it was in the training dataset. Important characteristics of VAEs can be listed as below:

1- VAEs produce a lower-dimensional representation of the input data (like classical AEs).
2- By design, VAEs have continuous latent spaces, which makes random sampling and also interpolation easier [106].

2.7 Generative Adversarial Network (GAN)

GANs are a set of generative networks, which are able to generate new content. In Generative Adversarial Networks (GANs), the aim is to produce new data (such as images) which look real. In fact, this goal is a min-max game between a Generator (G) and a Discriminator (D), so that D tries to recognize real and unreal images, while G tries to produce images that look real. This learning architecture gives these networks a good ability, suitable for frame processing tasks, some of which are listed below:

1- GANs allow CNN to learn an implicit distribution from data patterns [84].
2- GAN has good applicability for video prediction [64].
3- GANs are used to produce data for prediction applications, in which not enough training data is available [101].
4- GANs produce sharper images compared to VAEs [111].

2.7.1 Main challenges of GANs

Despite GANs special abilities in feature extraction and frame processing, there are some challenges, specific to these networks, which sometimes lead to a reduction in their use. The most noticeable challenges are:

1- They require a precise selection of hyper-parameters.
2- They need multiple initializations [23].
3- It is difficult to train adversarial methods, such as GANs.

2.8 U-Net

Autoencoders suffer from vanishing gradients and a lack of information symmetry in their architectures. To tackle this problem, U-Net is proposed, which adds a shortcut between a high-level layer and a low-level layer with the same resolution [99]. The difference between U-Net and AE, in architecture, is illustrated in Fig. 5.

2.9 Transformers

The attention mechanism has enhanced the performance of deep neural networks in various applications, including natural language processing and machine translation [34, 71], segmentation [66, 74], and other translation-related tasks. Drawing inspiration from the capabilities and successes of attention mechanisms, the Transformer architecture was introduced. A transformer is a deep learning based model that adopts the mechanism of self-attention to solve sequence-to-sequence tasks (e.g., sequence transduction, or neural machine translation) while handling long-range dependencies with ease. This model was primarily proposed in [121] for machine translation. This model later was extended to computer vision applications (e.g., vision transformer [32].). This model is composed of two parts (Encoder, Decoder); however, it is different from previous RNN-based sequence-to-sequence models. One main difference is that the input sequence can be passed parallelly to GPU so that the speed of training and inference can be increased. This model is also based on the multi-headed attention layer and overcomes the vanishing gradient issue easily. Transformers have recently gained more interest in video anomaly detection. Some of the related noticeable works are [59, 133].

Accurate video anomaly detection necessitates effective and precise modeling of both appearance and motion patterns. In this section, we thoroughly analyzed various neural networks, evaluating their capabilities in extracting and modeling spatio-temporal features. Additionally, we conducted comparisons between different networks, considering factors such as feature extraction effectiveness, speed, computational cost, and more. The topics covered in this section aim to assist researchers in selecting the most suitable network for their specific method and goals. Furthermore, they provide valuable insights into potential challenges and shortcomings that researchers may encounter in their methods.

3 Anomaly detection methods

This section provides a comprehensive review of anomaly detection methods from two distinct perspectives. Firstly, the methods are examined based on their approach to jointly

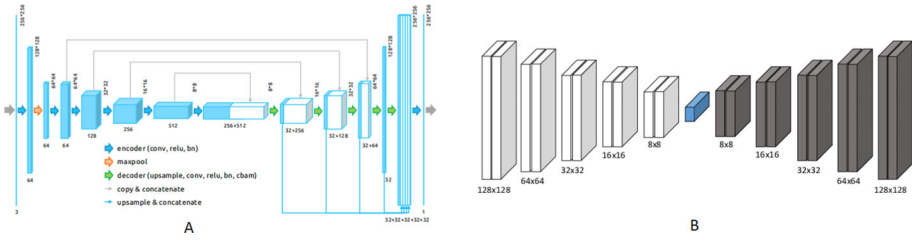

Fig. 5 Illustration of similarity and difference, in architectures, between U-Net (A) and Conv-AE (B). Figures A and B originally appeared in [1] and [46], respectively

extracting spatiotemporal features. This analysis offers researchers a comprehensive understanding of the architectural design of these methods, specifically how they model both appearance and motion features.

Subsequently, different methods are studied with a focus on how they formulate the task and approach the problem. This formulation is based on the proxy task they employ for modeling normal patterns. By reviewing these approaches, the capability and limitations of each proxy task in capturing the necessary features (such as object class, motion, color, etc.) for video anomaly detection are highlighted.

3.1 Methods based on Spatio-temporal (ST) feature extraction

A video is a sequence of frames, which are evolved over time. Therefore, the two main important defining attributes are appearance and motion, from which video analysis is performed. Appearance is the first attribute that attracts the attention of the analyzers. Anomalies in videos can be due to the presence of unknown (previously unseen) objects, which can be defined by appearance-based features. However, this is not the only factor for the definition or creation of anomalies in videos. In a variety of cases, it is the motion, which defines the anomaly. For example, an irregular speed of a car inside a street can determine an anomaly taking place in that scene. Similar to appearance, motion features should be analyzed and modeled both locally and globally, in order to gain a better understanding of video content. Motion patterns can be represented to the network directly by motion-based features (such as optical flow) or they can be captured by sequence-aware networks (such as the RNNs). The importance of considering motion is that most of the anomalies, in the real world, take place with moving objects. Humans consider and analyze both appearance and motion factors jointly and interactively, because, generally, motion and appearance are not always independent, but each one can also be a support to determine the other. For example, the motion pattern of an object (let us assume a snake here) can be a support for its recognition, in addition to its appearance features (such as shape, color, etc.). In video anomaly detection, understanding the Spatiotemporal context of a video is essential as it provides information about the evolution of the appearance of an event in a video [59]. Various methods and models are proposed for spatiotemporal feature extraction, which were studied in the previous section (literature on deep learning based models). However, from another viewpoint (the process of jointly extracting motion and appearance features), methods can be categorized into the following categories:

A- Single network-Single path methods: in this category (e.g., [63, 102, 115]), spatiotemporal features are extracted through a single path (single branch) process, by a single model. The most noticeable type of model in this category is C3D, which is able to extract rich spatiotemporal features for different applications, especially action recognition. However, a few important points should be considered regarding this model, as listed below:

  – C3Ds are difficult to train (high computational cost) and require an enormous amount of training data [21].
  – C3Ds capture local motion patterns [10]. C3Ds demonstrate effectiveness in modeling short-term motions. However, they encounter challenges when it comes to modeling long-term motion patterns, which is crucial in the context of video anomaly detection.

B- Two stream methods: In these methods (e.g., [110, 130]), motion and appearance are modeled separately, using two separate but usually identical branches. Generally, the input of

one of the branches is a raw frame and this branch is in charge of modeling appearance, through a frame reconstruction, prediction task, etc. In a complementary manner, the second branch attempts to capture and model motion patterns. This is generally achieved by receiving an explicitly-extracted motion feature (e.g., optical flow map) and modeling it through a reconstruction task [128], or by getting a raw frame and learning the associated motion patterns by predicting its corresponding optical flow map or optical flow magnitude map [11, 12] (i.e., through an image translation task). The two branches of the model are usually optimized jointly, which implicitly encourages the model to learn the features of both types in an integrated way. To better integrate motion and appearance, some methods such as [38], propose to add cross-branch connections, to transfer more information between the branches. Nguyen et al. [84] use a similar strategy, to jointly extract Spatio-Temporal features. In their approach, two identical but separate branches (i.e, decoders, in this example) decode the extracted features of a frame (produced by a common encoder), consecutively to reconstruct the input frame, and to estimate the optical flow. In the inference stage, they compute the reconstruction/prediction error of each branch in order to detect anomalies.

C- Hybrid methods: in methods of this category (such as [24, 107, 122]), generally multiple networks (each specialized in extracting specific features, such as motion and appearance) are connected in order to extract spatiotemporal features. As numerous research studies have proved, CNNs are powerful in image analysis and, on the other hand, RNN families are, by nature, suitable for analyzing video sequences. Hence several methods with different architectures have connected these two types of networks to extract suitable spatiotemporal features for anomaly detection.

It is worth mentioning that several fusion approaches, in several levels such as pixel-level (concatenation before ST feature extraction), feature-level (fusion of features before decision making), and score-level (fusion of anomaly scores extracted from different features) have been proposed to combine effects of different features, for a better anomaly detection [128]. The selection of the fusion type can impact the method's speed, complexity, and performance due to its unique characteristics. For instance, employing input-level and feature-level fusion can decrease the overall feature size, but integrating features of varying types and dimensions may present challenges [62]. Conversely, score-level fusion circumvents feature incompatibilities, albeit it may introduce high computational complexity and storage load [61].

3.2 Common approach of DL-based semi-supervised VAD methods

Various DL-based semi-supervised VAD methods have been proposed, which use different strategies for anomaly detection. They have approached the problem by performing various proxy-tasks (reconstruction, segmentation, prediction, etc.), which although are not of direct interest [87] and seem to be apparently unrelated to the task of anomaly detection, are very effective in video anomaly detection. All of these different methods, as a self-supervised task, mine different features [35], however, they all exploit the fact that all machine learning (ML) methods generally achieve the desired results for the data types on which they are trained (or at least for similar enough data types). In other words, ML methods do not guarantee to produce the desired results for test data, which is new and different from the training set. In the following sections, the methods are all based on the same general approach: they train a DNN (or any DL-based approach), on only normal data (the rule of semi-supervised methods), to perform a specific task. Hence, they produce the desired result for the normal

test data (that is, they detect no anomaly), since they have been previously seen during the training. The results would not be as desired for abnormal test data (that is, they detect an anomaly). The main challenge is to specify the desired scope of the task and select training and testing data accordingly. This way, DNNs would learn and use proper features of the video data and thus discriminate well between normals and anomalies at the test time. The following sections review important methods using this common approach, describing their strategies, their feature extraction procedure, and their strengths and weaknesses.

3.3 Reconstruction-based methods

In reconstruction-based AD methods, it is assumed that the models, trained by normal data, are able to reconstruct the normal test data accurately (i.e., with low reconstruction error), while the reconstruction error would be comparatively high for abnormal test data, which has not been observed by the model, during the training [44]. This methodology can be implemented in different ways. In various research studies, deep Auto-encoder networks (especially Conv-AE) have been used to learn to reconstruct normal data. AEs perform well in the reconstruction of the data, on which they have been trained. They encode the input visual data (a single frame, or a sequence of frames) into the latent space through an encoder and reconstruct the input data through a decoding pathway. For anomalies (the data samples, not seen in training), it is expected that the reconstruction error would be comparatively high [44]. An anomaly score (or vice versa, a regularity score) is normally calculated from the reconstruction error to indicate the anomalies. Figure 6 illustrates the process of reconstruction-based video anomaly detection.

One of the first and most noticeable works in this field is proposed by Hasan et al. [44], which uses a Conv-AE to extract spatiotemporal features from video clips and calculate the anomaly score from the reconstruction error. Although it uses a group of consecutive frames

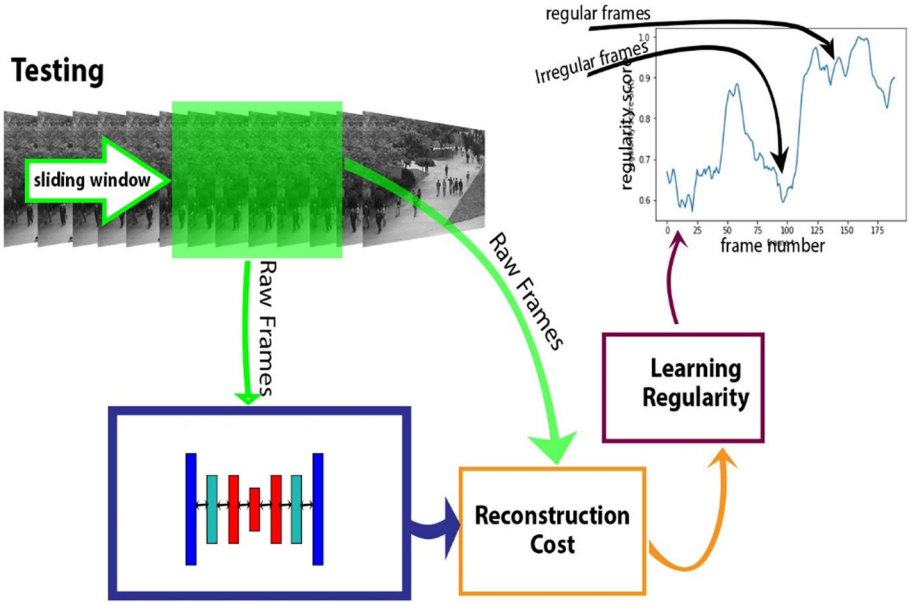

Fig. 6 Reconstruction-based video anomaly detection, using AEs. This figure originally appeared in [104]

as input, instead of a single frame, to enforce the model to capture temporal dependencies, the 2D convolution destroys the temporal information, after the first convolution layer [118]. This issue has been addressed in [24], which proposes a Conv-LSTM-AE to learn spatiotemporal features. The model extracts the spatial features of the frames, by a Conv-encoder, and passes them to a LSTM encoder to track temporal variations; the output goes through a reverse (temporal and spatial) decoder to reconstruct the frames group. Conv-LSTM-AE has also been used in [70] for anomaly detection. In other similar works such as [23] and [137] a similar approach has been used for anomaly detection; however, they model normal data by minimizing the difference between the latent spaces of the input frame and the reconstructed frame, in addition to minimizing the reconstruction error of the frame itself. The work in [33] proposed to reconstruct the optical flow map of each frame, in order to consider the motion and to detect the anomalies in video. Moreover, some other researchers have proposed the concatenation of the appearance (frame) and motion data (optical flow), as an input, for the purpose of reconstruction. Nguyen and Meunier [84] proposed to use two different branches for motion and appearance, in order to capture the correspondence between them and to detect the anomalies more effectively. In this way, one branch is in charge of frame reconstruction (capturing spatial dependency) and the other one attempts to estimate the optical flow map, to capture motion dependency, customized for the task. Different models that have been applied for representation learning or reconstruction-based anomaly detection are as follows: PCA [72], classic AE [44], Conv-AE [44], Contractive-AE [96], Conv-LSTM-AE [24, 77, 122], Hybrid Spatio-Temporal Autoencoder [125], Denoising AEs [82] and VAE [127], GRU-AE [81]. Some of the other examples in this field are [3, 45, 69, 80, 97, 128]. Manassés et al. [76] study the deep convolutional auto-encoders for anomaly detection in videos.

3.3.1 Challenges of auto-encoders in reconstruction-based VAD

As can be noticed from the references mentioned above, AEs are the main tool for reconstruction-based video anomaly detection methods. Below, some challenges and shortcomings of AEs for VAD are listed.

1: AEs have a high learning capacity and a good power of generalization. Hence, the assumption that anomalies have a high reconstruction error is not always true [64, 136]. Researchers in [59, 68] believe that this generalization issue is because AEs learn an identity map between inputs and outputs. These works leveraged prediction proxy-task for anomaly detection to handle the mentioned challenge.
2: When an AE is trained to minimize the Mean Square Error (MSE) for frame reconstruction, the network actually learns the average of previously seen training data.
3: Anomalies occurring in small regions can be neglected, because of the adding and averaging process for the entire frame, which may produce a low reconstruction error for anomalies in small regions. Nguyen and Meunier [84] proposes using small patches instead of the entire frame in its score estimation scheme, to partially handle this challenge.

3.4 Prediction-based methods

Prediction, generally, means the estimation of the masked frame(s), based on previously seen frames. Prediction-based VAD methods cast anomalous events as unexpected events in future frames [120]. In prediction-based anomaly detection methods, it is assumed that predictive models, which are trained on normal sequences (previously seen frames), can

precisely predict the masked frame(s) (usually future frames) in normal test sequences but their prediction error would be comparatively high in abnormal test sequences. Thus, in video anomaly detection, video frames are considered as sequences and the goal is learning the normal patterns (appearance and motion), in consecutive normal frames and predicting the masked frames, based on the learned patterns. The prediction error can be easily calculated by measuring the difference between real and predicted frames or by calculating the conditional probability of a new observation based on the previous samples [55]. Different constraints have been used for anomaly detection, in prediction-based frameworks, such as appearance (gradient and intensity) and motion [64]. Experiments in [64] show that predicted frames for abnormal samples are unclear and usually with color distortion and it is claimed that among several networks, GANs show better results for video prediction. Wang et al. [123] report that feature extraction, through the prediction process, has high quality and it is more suitable for video analyzing applications since accurate prediction highly depends on high-quality features. It is worth mentioning that, in prediction-based methods, the input and output are not necessarily of the same type or size and they can be different, in different approaches. For example, [100] takes advantage of two cross-domain generators, in which one learns to predict the past gradients from appearance and the other learns the reverse, for local anomaly detection. Prediction-based video anomaly detection strategy has been utilized in numerous research studies such as: [22, 64, 67, 68, 94, 107, 131].

3.4.1 Generative models for reconstruction and prediction

Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAEs) are also widely used in the reconstruction and especially prediction-based anomaly detection methods. The main difference between these methods and previously introduced approaches is that these methods consider the distribution similarity, in addition to pixel-wise similarity. Some of the noticeable AD works, based on GANs, include the researches conducted by Zenati et al. [135], Kimura et al. [54], Akcay et al. [6], Ravanbakhsh et al. [93], Akcay et al. [7], Sabokrou et al. [103], Gherbi et al. [41] and Ganokratanaa et al. [39]. Zenati et al. [47] and Donahue et al. [30] use a biGAN to map a latent space to an image and use it for anomaly detection. Gans have promoted the performance for various AD approaches, especially prediction-based approaches. However, as GANs may show instability during training, their usage for anomaly detection may be limited. Hence, in order to address this problem, in addition to comparing frames, extracted features are also compared to calculate the loss [6]. Galeone et al. [28], Rani, and Sumathi [15] have, comprehensively, studied GANs for anomaly detection.

3.4.2 Prediction versus reconstruction

A prediction-based method attempts to obtain the most information from the most recent frames, as they are more relevant to the future frame [77, 78]. Hence, predictive methods lose a lot of information about the past and their generic (general) prediction would be less precise. Moreover, Pathak et al. [87] declare that as nearby frames are visually similar (considering the texture and the color), they might focus on learning low-level features instead of high-level semantic features. Reconstruction, on the other hand, attempts to learn an obvious representation from data [77] and in fact, it memorizes the input [55] and considers all frames almost equally. In this way, it neglects the temporal evaluation between frames. To address the mentioned challenges, a composite approach has been proposed to benefit from the advantages of both methods [81, 113, 132]. For example, the proposed LSTM-AE network

in [113] (Fig. 7) is composed of two branches, one for reconstruction and one for prediction. These branches have an encoder in common but with two separate decoders. Some other examples of composite VAD methods are [65, 78].

One of the challenging aspects of both prediction-based or reconstruction-based methods is that even slight lighting variations may cause a high pixel-based loss, which can be deceptive. Moreover, these approaches generally train the model (reconstruction or prediction) from scratch, in an unsupervised manner, and the entire frame or only proposal patches are reconstructed or predicted. Hence, these approaches are not aware of the class of the objects in the frames. To address these challenges, researchers such as Bergmann et al. [14] use a pre-trained network (trained on natural images) as the encoder. Producing the latent space in this way, helps the network leverage prior knowledge about the nature of the natural images and tackle the issue to some extent.

3.5 Memorization-based methods

One of the main challenges with previous methods is that DNNs (and especially CNNs) are so powerful in generalization, that they may reconstruct the abnormal frames too well. Hence, the assumption that the reconstruction/prediction error is comparatively high for

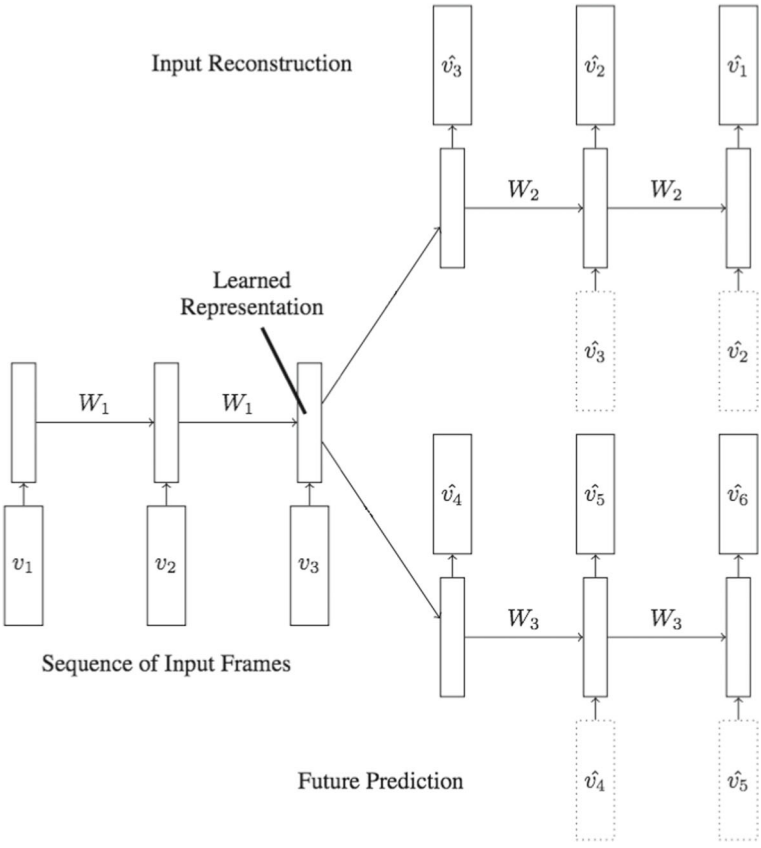

Fig. 7 Combination of frame reconstruction and prediction for anomaly detection in video [113]

abnormal test frames is not always true. In order to address this problem and reduce the representation power of DNNs, memorization-based anomaly detection methods have been proposed. These methods use the encoding of the input frame as a query to select the most relevant saved items, from the recorded prototypical patterns of normal data, to reconstruct the input frame. Consequently, the previously recorded items are decoded and selected from memory, instead of using the output of the encoder directly. For example, Gong et al. [43] proposed MemAE (Memory augmented AutoEncoder) which learns and updates the memory contents, during training, to represent the prototypical elements of the normal data. In the test phase, the memory is fixed and reconstruction is performed using items selected from the memory (see Fig. 8). Moreover, Park et al. [85] propose a similar strategy for anomaly detection and reconstruct or predict a video frame with a combination of items in the memory, rather than using CNN features directly from an encoder. In this work, items in the memory record prototypical patterns of normal data and the diversity of normal patterns is considered explicitly, since the authors believe that a single prototypical feature is not enough to represent various patterns of normal data. Other examples of memory-based approaches can be found in [65, 107, 122].

3.6 Object-centric based methods

As mentioned before, one of the main shortcomings of the methods based on frame reconstruction or prediction is that they do not explicitly (and hence effectively) consider the objects. Object-centric approaches concentrate on detected objects (detected by state-of-the-art object detectors) and study their appearance and motion features to make decisions. The research conducted by Ionescu et al. [48] is one of the recent works on video anomaly detection that detects objects of interest to accomplish anomaly detection. Moreover, Doshi and Yilmaz [31] propose an object-centric approach, in which objects of interest in each frame are detected by a pre-trained YOLOv3 object detector and consequently, a feature vector containing appearance, motion, and location information is extracted to learn normal behaviors. Unlike Ionescu et al. [48], this method considers location information by containing a summary of location information in its provided feature vector. Other researchers have also detected anomalies by detection of objects [16, 50, 109, 126]. The advantages and challenges of object-centric based methods are:

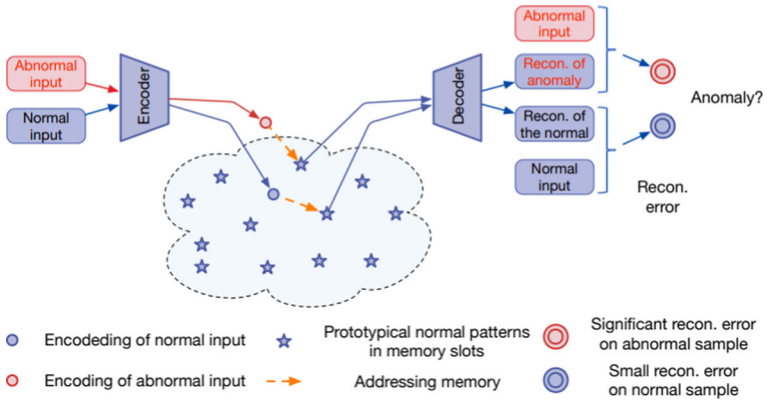

Fig. 8 Illustration of the memorization-based anomaly detection [43]

– The objects are explicitly considered, which is helpful in video understanding.
– The anomalies are easily located inside the frame.
– The performance of the method completely relies on the object detection part.
– Information regarding the object size and the context (such as the location information) is removed as these methods crop and resize the detected objects (Fig. 9).

3.7 Segmentation-based methods

Krzysztof et al. [56] perform the anomaly detection task in a different way. The proposed idea arises from the fact that a semantic segmentation approach can segment the objects properly if it has observed them in the training phase and it would show worse results for unseen objects. This fact can be used for image anomaly detection. The researchers propose to synthesize the image from produced semantic segmentation maps and the reconstructed images help to define and locate the novel objects. The positive point of this method is object-type awareness, however, this method is proposed for images, not videos. Another research [26] proposes a similar anomaly detection approach, based on foreground segmentation and detects unexpected objects (i.e., objects not seen in the training samples). To extend this idea to video anomaly detection, inspired by [56], Mohammad et al. proposed a two-stream segmentation-based VAD method [11], which leverages knowledge distillation to propose an object-class aware video anomaly detection method. In the appearance stream of their proposed method, a teacher-student strategy (Mask-RCNN as teacher and resnet-UNet as student) is proposed. The student network learns semantic segmentation with the annotations generated by the teacher for each input frame. In the inference stage, the trained student fails to semantically segment abnormal objects precisely and the error of segmentation is used to calculate the anomaly score. Mohammad et al. also proposed a similar method in

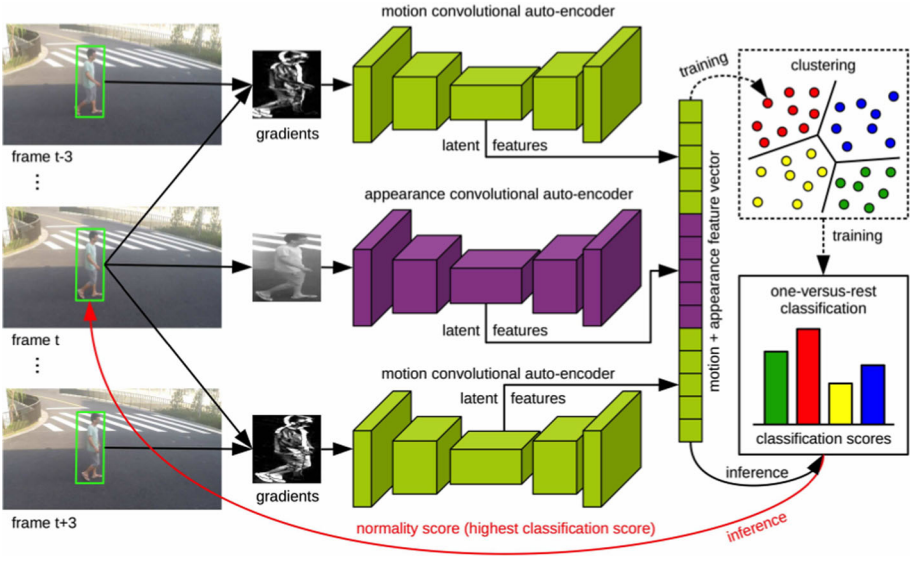

Fig. 9 An object-centric video anomaly detection method, proposed in [48]

[12] to predict the semantic segmentation of the future frames by observing the previous two consecutive frames. This helps the network detect sudden motions in addition to detecting unexpected objects via a single task (i.e., predicting the semantic segmentation map of the future frame) (Fig. 10).

3.8 Multi-task learning based AD methods

Various proxy-tasks have been proposed by researchers for anomaly detection. However, approaching anomaly detection problem through a single proxy task is suboptimal since a single proxy-task is not able to detect all anomalies and is not well aligned with the problem [40]. Hence, researchers have recently proposed to use multiple proxy tasks in one method to cover more anomalies. One of the outstanding multi-task learning based VAD methods is proposed by Georgescu et al. in [40]. This method benefits from 4 self-supervised proxy-tasks such as (i)discrimination of forward/backward motions, (ii)discrimination of objects in consecutive/intermittent frames, (iii)object-specific appearance reconstruction, and (vi)knowledge distillation tasks for effective anomaly detection. Mohammad et al. [12] proposed another multi-task learning based method with different proxy tasks. In this work, they combine the abilities of future frame prediction task and semantic segmentation task to the novel task of future semantic segmentation map prediction, to detect appearance and motion anomalies via a single task. He also proposes a knowledge distillation task in another stream, to detect other types of motion anomalies. In this task, the student learns to map input video frames to their corresponding optical flow magnitude maps (produced by the optical flow extractor method as a teacher). In this formulation, the network learns to associate each object with its normal motion and finds the anomalies by calculating the differences between the estimations of the teacher and the student network. Generally, it is assumed that adding more proxy-tasks

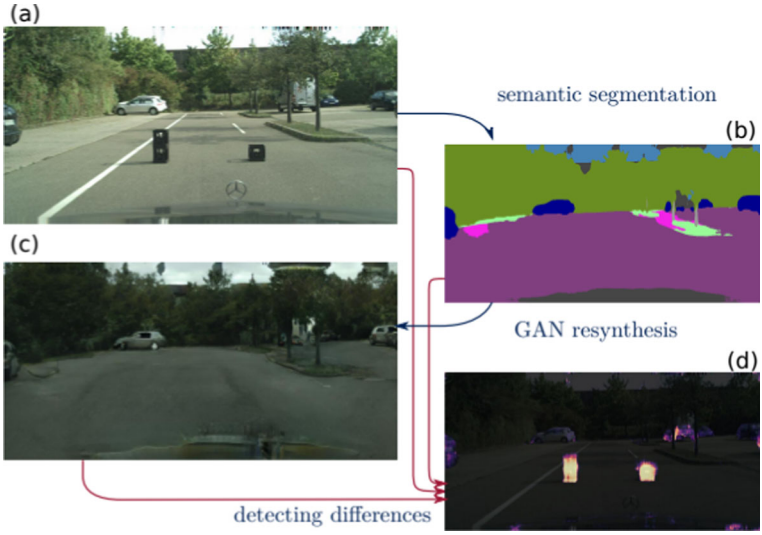

Fig. 10 Anomaly detection, based on semantic segmentation [56]. a: input frame. b: extracted semantic segmentation map. c: resynthesized frame from segmentation map. d: difference of input and resynthesized frame. The image in this example comes from the Lost and Found dataset [89]

could results in the detection of more anomaly types, however, adding more tasks adds to the complexity of the method, run time, memory consumption, etc. Hence, the key idea in multi-task learning based methods is to propose the least number of complementary tasks to cover more anomalies but with fewer parameters. Another example in this field is presented in [20, 59].

3.9 Shortcomings or challenges of previous methods

Through a meticulous analysis conducted in the previous sections, numerous deep learning-based semi-supervised video anomaly detection methods were subjected to critical evaluation. Drawing insights from the experiments and conclusions of these methods, several challenges and shortcomings have come to light. These findings serve as valuable cues for researchers in the field, providing valuable guidance for future endeavors and shedding light on areas that require further exploration and improvement. The following summary highlights these significant challenges and shortcomings, presenting them as vital considerations for researchers in their future works:

1: Previous holistic methods usually reconstruct/predict the entire frame/map and consider the appearance and the motion with a global look. They do not consider the objects and other details individually. Hence, the performance of these methods can be affected by background errors. Object-centric approaches on the other hand, only attempt to focus on reconstruction or prediction of the objects and therefore fail to consider the context. Hence, almost all existing methods neglect some important information in their algorithm.
2: A considerable portion of spatiotemporal information in frames is redundant, and is not required in scene analyzing or video understanding. This leads the network to divide its attention to a variety of aspects (including these redundant parts) and not to precisely focus on useful portions (for example, to the objects of interest). This fact plays an important role in video anomaly detection since anomalies generally occur rarely and may occupy a small portion of a frame. This problem has not been acceptably covered in existing methods.
3: Loss functions, which direct the network to capture effective features, do not simultaneously and effectively apply compactness and descriptiveness constraints to the feature extraction process.
4: The relation between the class of the object of interest, its motion, and its location has not been taken into account effectively in existing methods.
5: In the existing methods, if an anomaly occupies a small portion of the frame, its effect could be lost on the anomaly score of the frame. On the other hand, even object-centric methods may also have a difficult time detecting such small objects.
6: The currently used approach for the calculation of the reconstruction error is not reliable. Any small changes in all pixels (for example illumination changes in the environment) can result in a high change in reconstruction error.
7: Holistic models, trained on a scene, may not perform well after a change in the scene or viewpoint.
8: In existing methods, the fusion of different anomaly scores (e.g., motion anomaly score, appearance anomaly score, etc.) does not apply the effect of all factors effectively and one factor may dominate others and lead to underestimation of other factors.

4 Experiments

The preceding section offered an in-depth review and evaluation of various methods, drawing from the outcomes of past research studies. These analyses would also be drawn theoretically, taking into account various aspects such as the structure of the proposed network, the target or loss function, and the outcomes from analogous applications, among other factors. The alignment between these theoretical and empirical findings fosters the notion that drawing upon the knowledge gleaned from DNNs (as discussed in Section 2) and the empirical outcomes of prior methods (covered in Section 3), additional theoretical hypotheses or conclusions can be formulated. These can then be validated through new experimental studies. Therefore, in this section, we carry out a series of experiments to confirm some theoretical conclusions that are not explored in earlier studies. The key points to be investigated are as follows:

- The effect of the number of foreground objects: For any proxy-task selected for anomaly detection, an anomaly score is calculated based on how well the trained model performs on normal and abnormal data. The underlying presumption is that the model's performance is significantly inferior when dealing with anomalies compared to normal instances.For instance, the reconstruction error is anticipated to be greater for anomalies. However, even though the model is trained on normal data, achieving zero reconstruction error is practically unattainable for normal instances as well. Put differently, even though the anomaly map is expected to exhibit higher activations for anomalies, there will also be some activations present for normal regions. Therefore, we can infer that in holistic approaches (i.e., methods that take into account the entire frame for anomaly detection), the anomaly score of a frame also relies on the number of objects within that frame. This could lead to a higher anomaly score for a frame filled with numerous normal objects, as compared to a frame containing a single abnormal object.
- The effect of the camera distance (or the object size): The qualitative results from all prior methods indicate that the activations in the anomaly map for each object (which could also be considered as the model's uncertainty for that object) are dependent on the size of the object. Therefore, it can be inferred that the anomaly score of a frame for larger objects (or objects nearer to the camera) would be higher compared to frames with similarly typed but smaller-sized objects. This issue can lead to a higher anomaly score for a larger normal object compared to a small abnormal object.
- Awareness of the method concerning the class of the objects: The proxy tasks associated with frame reconstruction or prediction, along with their loss functions, are designed to incorporate low-level image features (like intensity, color, etc.) for anomaly detection. The models used in these methods are primarily focused on learning these low-level features in order to accurately reconstruct or predict the desired frames along with their image specifics. This arrangement does not ensure that the model takes into account object class information to execute the proxy task. Consequently, these methods emphasize the detection of novelties or deviations in low-level features (such as color) and may assign a higher anomaly score to a normal object with unusual colors than to an abnormal object with a color frequently encountered during training.
- Illumination changes: As mentioned in the previous point, formulating anomaly detection using proxy tasks that primarily focus on low-level features can make the method susceptible to changes in illumination within the frame. Therefore, we can anticipate a higher anomaly score for frames with different lighting conditions than what was encountered during the training phase.

– Effect of motion patterns: Experiments from past studies demonstrate that CNN-based models tend to prioritize appearance features over motion features, dedicating more effort to learning appearance characteristics as compared to motion. Therefore, video anomaly detection methods, which are designed to detect both appearance and motion features through a single branch, might lean towards and prioritize appearance features, potentially leading to the neglect of motion information.

To analyze these points, we implemented two state-of-the-art methods. Our analysis will yield insights applicable not only to these particular methods but also to others that share similar components or detection strategies. We approached these experiments from a unique perspective: examining the instances where these methods fail. Rather than comparing the numerical performance of various methods, our aim is to illuminate the respective strengths and weaknesses of different approaches. This will help future work to address these areas, ultimately reducing the occurrence of false positives and false negatives.

First, we implemented the method proposed by Hasan et al. [44] which uses a Conv-Autoencoder for anomaly detection. It is worth mentioning that Hasan et al. conducted experiments on two different autoencoder architectures: Fully Connected Autoencoder (FC-AE) and Convolutional Autoencoder (Conv-AE). In addition to the experimental results that they provided, it also can be concluded (considering the network architecture) that FC-AE destroys the structure of the image and it could not show comparatively promising results for image processing. Hence, we do not implement that part. To train the Conv-AE, we implemented the same architecture and used the same hyperparameters as originally proposed (The implemented architecture is shown in Fig. 11).

After training the model on the UCSD dataset, we performed an evaluation on its test dataset. We calculated the reconstruction error and consequently the regularity score for each frame as in (1) and (2). In these equations, s(t) and e(t) show the regularity score and reconstruction error of the frame, respectively. I(x,y,t), e(x,y,t) also refers to the intensity and the reconstruction error of the pixel.

$$e(x, y, t) = \|I(x, y, t) - f_w(I(x, y, t))\|^2 \tag{1}$$

$$S(t) = 1 - \frac{e(t) - min_t e(t)}{max_t e(t)} \tag{2}$$

As can be seen in Fig. 12, the results of the experiments show that this method fails (generates false positives) when the number of foreground objects is considerably variable in different frames. As Fig. 12 shows, when the number of foreground objects is high, we

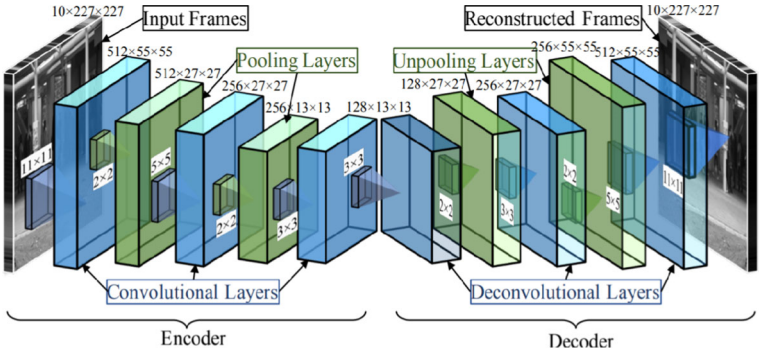

Fig. 11 The architecture of the Conv-Autoencoder proposed in [44] for video anomaly detection

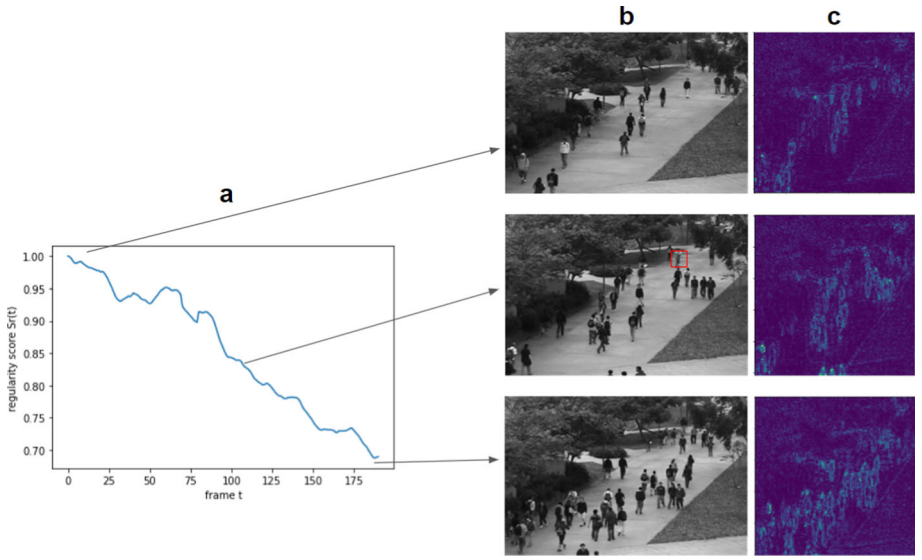

Fig. 12 Effect of the number of the FG objects on the regularity score (UCSD-Ped1-Test003). (a) Regularity score as a function of frame number. (b) Samples of frames in the test clip. (c) Reconstruction error map for the input frame (e(x,y,t))

would have a lower regularity score (or higher reconstruction error) for the frame. This is because each object, more or less, has a reconstruction error and as the number of the objects rises, the total reconstruction error of the frame, which is the sum of the errors of the foreground objects and the background (BG), increases. From another point of view, as Fig. 13 shows (This figure originally appeared in [44]), the most regular frame for each scene is an image quite similar to its BG. BG pixels are the constant and the most frequent pixels in all images during training and the network easily learns them. The reconstruction error of the frame is due to the difference between the input frame and the most regular frame (let us assume BG here) and is thus directly affected by the number of foreground objects. It can be concluded that, for cases in which the class of the objects defines the anomalies rather than their number, objects should be analyzed individually (as with object-centric approaches) instead of evaluating the entire frame at once. We also repeated the same analysis for the training samples (Fig. 14) and the experiment confirms the previous results. That means that the reconstruction error of the frames with high populations (even if they do not contain any

Fig. 13 (a) Synthesized regular frame for Ped1. (b) Synthesized regular frame for Ped2

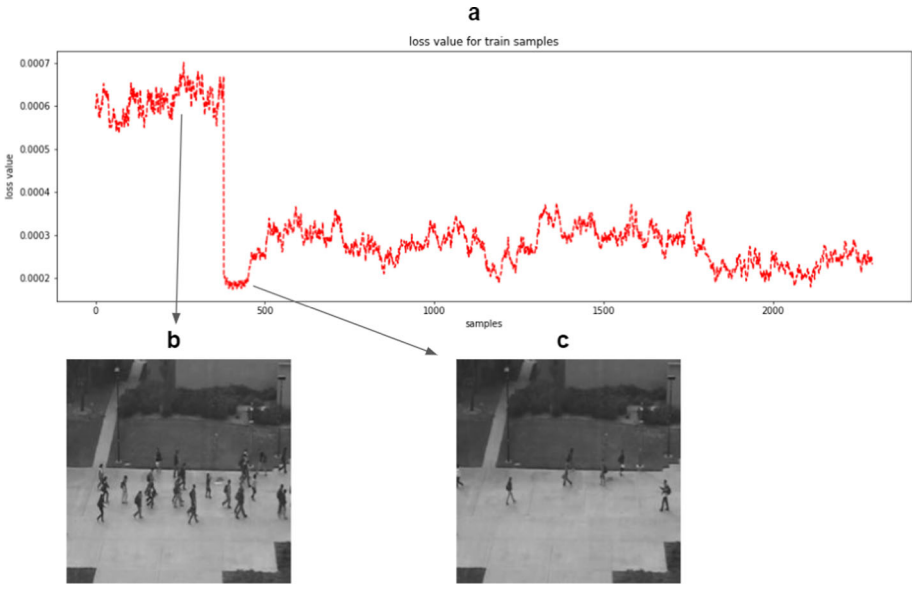

Fig. 14 Effect of the number of the foreground objects on the reconstruction error (results on Ped2). (a) Reconstruction error (loss value) for the training frames. (b) A frame with a high number of people. (c) A frame with a low number of people

anomaly) is considerably higher than that of the other frames (even compared to frames with anomalies).

In most of the recent research studies, the appearance and motion features were analyzed separately, in separate branches. In the motion branch, researchers reconstruct or predict the previously extracted motion features (such as optical flow or difference of consecutive frames) to learn the normal motion patterns. We analyzed the effect of the number of objects on the results of the motion branch. As can be seen in Fig. 15, the results give rise to the previous conclusion; if the frame (or the motion map) is analyzed globally (analyzing the entire frame, not each object) the reconstruction error would be affected by the number of foreground objects, rather than the important anomaly factors.

We should make this point clear that the mentioned shortcoming has nothing to do with formulating the anomaly detection as a reconstruction or a prediction problem but is due to the fact that these methods consider the frame holistically. As object-centric approaches analyze each object individually, they do not face this challenge.

Considering the frame entirely has a strong point: Considering location. Frame reconstruction or prediction-based methods, mostly learn a pixel-wise model. This means that they learn a model for each pixel separately, and pixels at different locations expect different intensities. Before examining the experiment results, let us have a second look at Fig. 13-b, for example. Figure 13-b shows that the most regular frame for the Ped2 has many dark pixels on the upper side of the walkway. This is because, during training, the network has frequently seen dark objects on that side. From this image, it is more expected (i.e., it is normal) to see dark objects on the upper side and the presence of dark objects in the lower part of the walkway most probably would be detected as an anomaly. To validate this idea, we analyzed the response of the network in some frames of Ped2, which is shown in Fig. 16. As expected, the network considers the position of the object in the scene and it produces higher

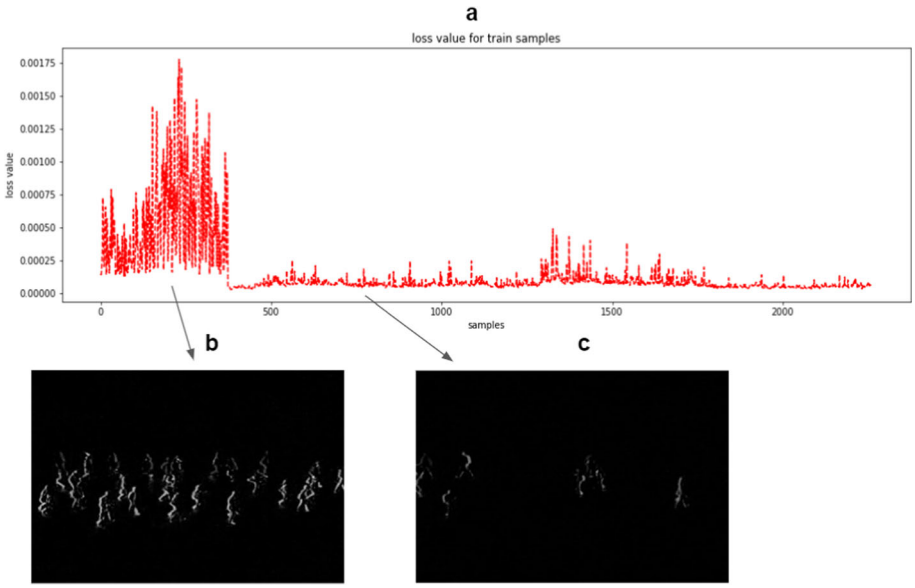

Fig. 15 Reconstruction error for motion maps. (a) Reconstruction error (loss) of motion maps for training samples. (b) Motion map of a frame with a high number of people. (c) Motion map of a frame with a low number of people

errors for objects which are not expected in that location. This positive feature is missed in object-centric approaches, as they crop objects out of the frame and analyze them individually. Thus, the location information is either lost or not considered. This would produce false positives and false negatives for object-centric approaches in different datasets, especially the street scene dataset. As can be seen in Fig. 1, in the Street scene dataset, the definition of normality (and hence anomaly) is different for 4 different cars, considering their locations.

Figure 17 shows the results of anomaly detection for different sample frames. This figure shows the input image (a), the reconstructed image by the model (b), the reconstruction

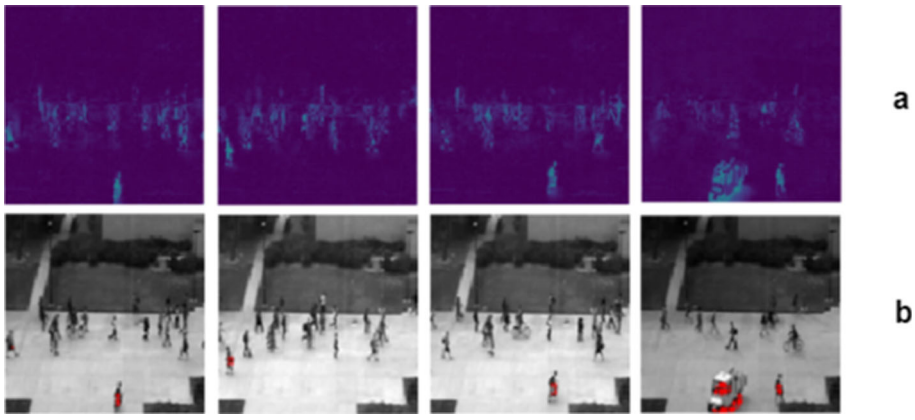

Fig. 16 Different responses of the network to the same pixel intensities at different locations. (a) Reconstruction error map (b) Input Frame with detected anomalies in red

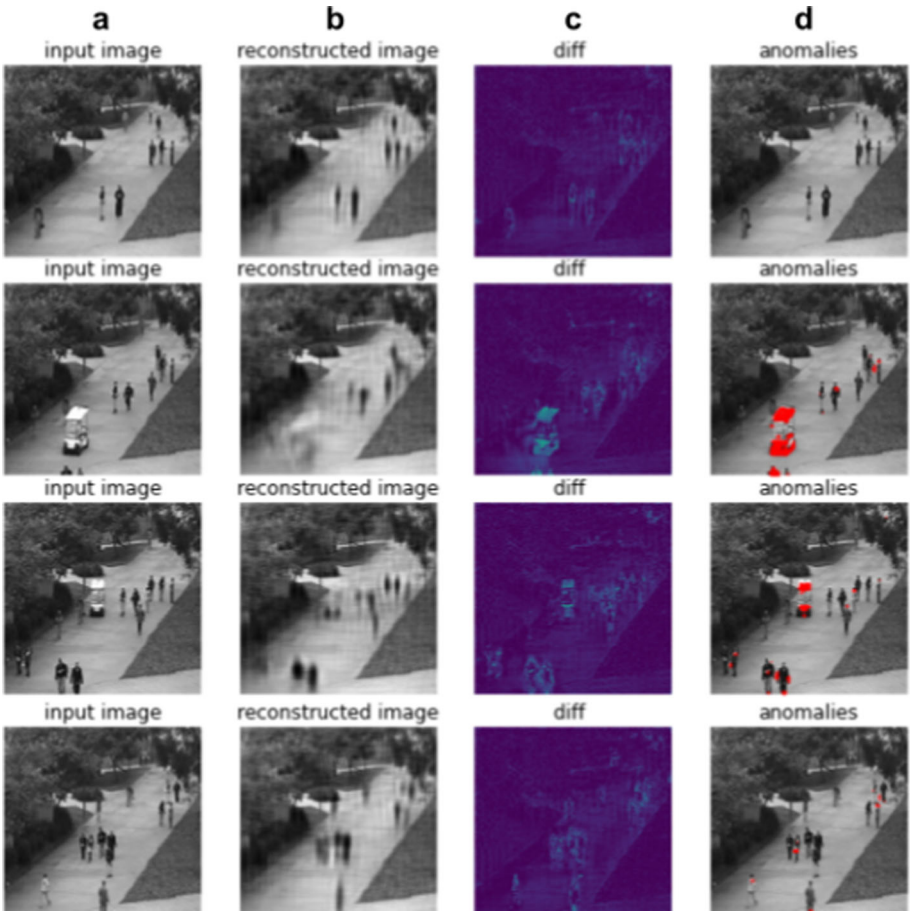

Fig. 17 Effect of the camera distance and pixel intensity on anomaly detection (results are provided for UCSD-Ped1). (a) Input frame. (b) Reconstructed frame. (c) Difference between input and reconstructed frame (reconstruction error map). (d) Result of anomaly detection (anomalous pixels are indicated in red)

error map or the differences between the original and reconstructed image (diff) in (c), and finally (d) indicates the pixels which are most probably anomalies, by applying a threshold on the reconstruction error map. This figure highlights a few points: First, what is missing in existing methods (in most reconstruction and prediction-based methods) is that the effect of the distance of the object from the camera is not considered. As illustrated in the figure, as the car approaches the camera, it occupies many pixels in the frame and produces a higher reconstruction error for the entire frame. Hence, the anomaly score of the frame would be affected by the factor of object distance. Another point that can be concluded from the results is that these methods mostly consider the intensity (or color) of the objects instead of the class of the objects. As illustrated in the results, the anomaly points (red points) are only detected for the pixels which have an intensity that differs from that of the background, and the other parts of the car are detected as normal. This can also be concluded logically; in these methods, the model is trained to reduce the Mean Square Error of the pixel's intensity (low-level features), which causes it to focus on low-level features. On the other hand, no information regarding the class of the objects is provided to the model directly.

The method proposed by Hasan et al. does not effectively consider motion, because as mentioned in Section 3.3, the first convolutional layer destroys the temporal information. Furthermore, our experiments do not show any considerable reconstruction error for objects with abnormal motion (faster motion here) such as bikers and skateboarders. Although in the results there is a noticeable reconstruction error for the cars, due to their different pixel intensity and their comparatively larger size. Hence, in the second step, we implemented the method proposed by Chong et al. [24]. This method benefits from a temporal autoencoder which is embedded inside the spatial autoencoder. We implemented the same architecture, as proposed originally [24]. This architecture is shown in Fig. 18.

The results indicate that the previous challenges in Conv-AE such as the effects of distance and number of objects, unawareness regarding the class of objects, etc. still exist here since this approach has similar strategies (such as evaluating the entire frame at once, focusing on intensity, etc.). However, as this method adds a temporal autoencoder to the network, it can capture motion patterns.

As illustrated in Fig. 19, bikers' bodies (unlike other persons' bodies) produce a higher intensity in error maps. The proposed method explicitly models the temporal evolution of the frames and hence can capture motion. However, the produced reconstruction error for the entire frame again depends on several factors which may degrade the effectiveness of the method. As it can be concluded from these results and also previous ones, these factors can be: 1) Number of foreground objects: the number of the objects is more decisive than the effect of the object motion. 2) Distance from the camera: the motion effect of an object, in the reconstruction error map, can be easily neglected if the object is located far from the camera (i.e., for smaller objects). 3) The results show that in single path methods (categories A and C in Section 3.1), the effect of the appearance features may dominate the effect of motion features. It can be expected that two-branch approaches would produce better results in considering the effect of the motion anomalies, as they independently consider and analyze motion in a different branch. Figure 20 confirms the previous point. In this experiment, we removed 9 consecutive frames (frames 9 to 17) to synthetically generate a sudden motion between frames 8 and 18. Through these frames, all objects inside the scene are normal objects. In other words, we synthetically generated an abnormal motion for normal objects (i.e., the abnormality is simply due to the motion factor), and this motion is much faster than

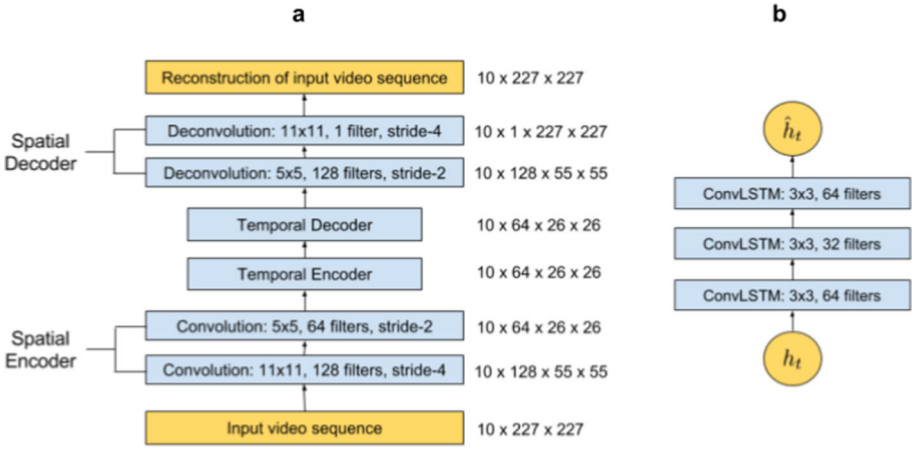

Fig. 18 The proposed Conv-Lstm-Autoencoder in [24]. (a) The entire architecture. (b) The temporal autoencoder

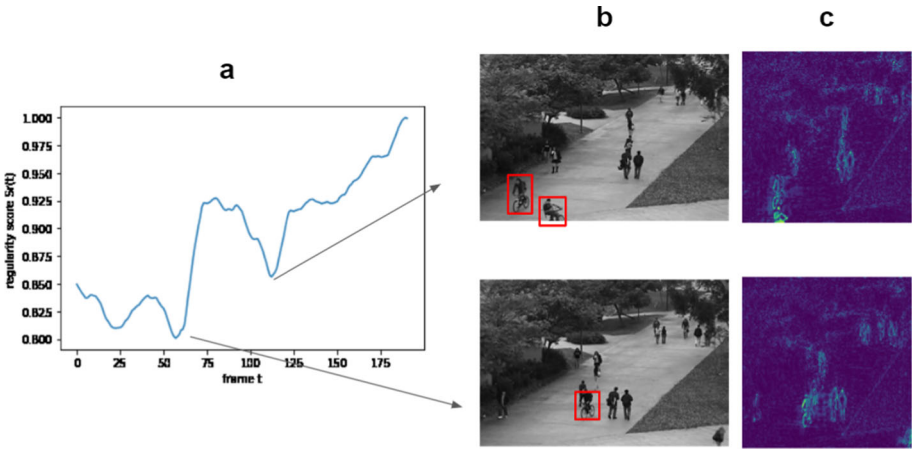

Fig. 19 Results of Conv-Lstm-AE on Ped1 (Test033). (a) Regularity score of the clip. (b) Input frames. (c) Reconstruction error map for the input frames

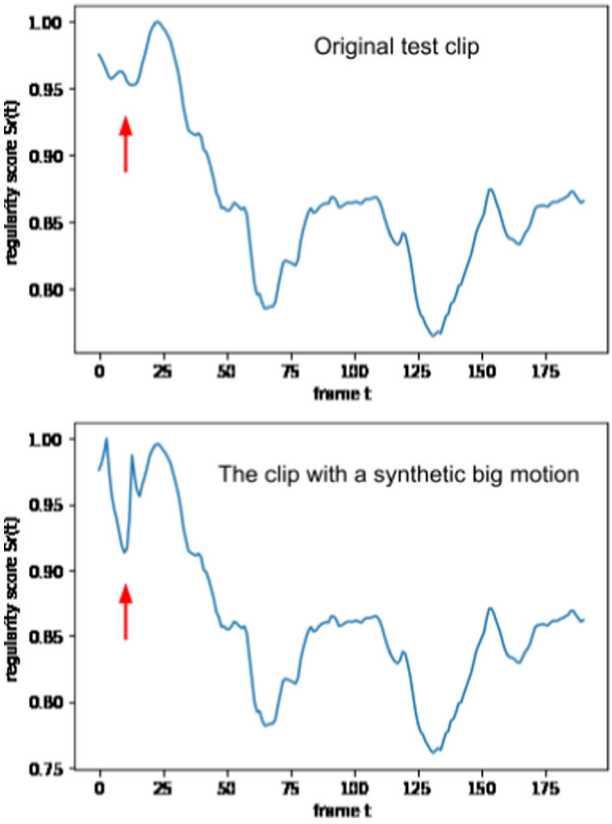

Fig. 20 Regularity score for different motions. (a) Regularity score for an original test clip (ped1-test016). (b) Regularity score for the same clip with synthesized large motion

any motion in the clip. However, as Fig. 20 indicates, the regularity score fall (anomaly score) is not considerable.

In another experiment, in order to analyze the effect of lighting changes in the scene we made changes in the brightness of some test frames. For this purpose, the pixel intensity of entire pixels in a test frame was multiplied by 1.3, and the reconstruction error map was extracted for the original test frame and the newly generated frame. The results, shown in Fig. 21, indicate that by lighting changes in the frame, the reconstruction error map varies considerably, which affects the performance of the method. In other words, this not only shows that the system is vulnerable to lighting changes but also that the system considers low-level features instead of focusing on high-level ones.

Experimental results in Figs. 22 and 23 illustrate the dominance of some factors such as intensity and distance on the class of the objects (which is the reason for the definition of the anomaly here). In Fig. 22, the top two frames are both normal frames, however, the second one would most probably be detected as an anomaly. In Fig. 23, the frame in row 4 contains an anomaly in the far distance which results in producing a higher regularity score compared to row 1 which only contains normal objects.

In the final step, we carried out a test to evaluate how effective object-centric approaches could be, in considering the class of the object and identifying abnormal objects. Object-centric approaches, as discussed in Section 3.6 and as observed in Fig. 9, crop objects out of the frame and train a network (usually an autoencoder) to learn normal patterns. However, as the experiments show, reconstructing a frame or even the appearance of objects individually, does not necessarily lead these approaches to consider the class of the object. This is mainly due to the fact that training an AE for the purpose of reconstruction, mostly focuses on the intensity (or color). What object-centric approaches mainly contribute, is focusing on the

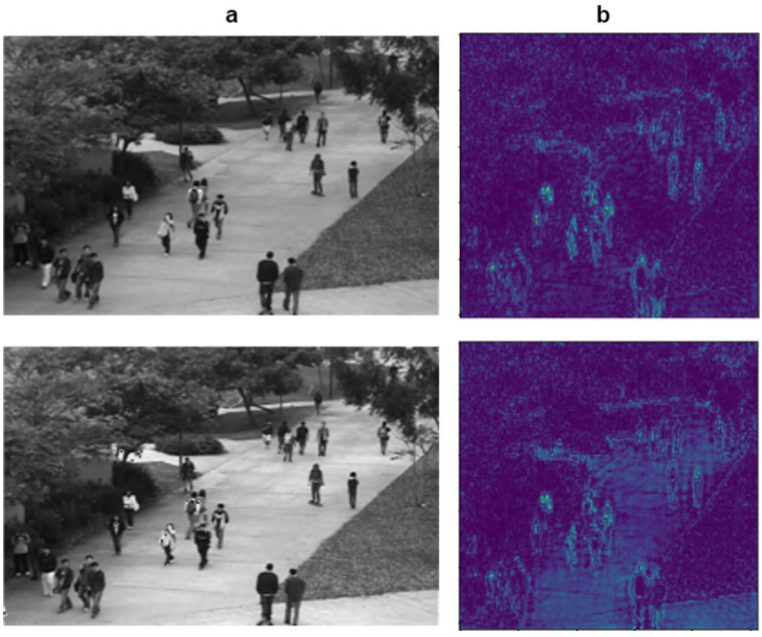

Fig. 21 Effect of lighting changes on the reconstruction error map. (a) Input frames. top: original frame, bottom: the same frame after increasing the pixel intensities. (b) Reconstruction error map

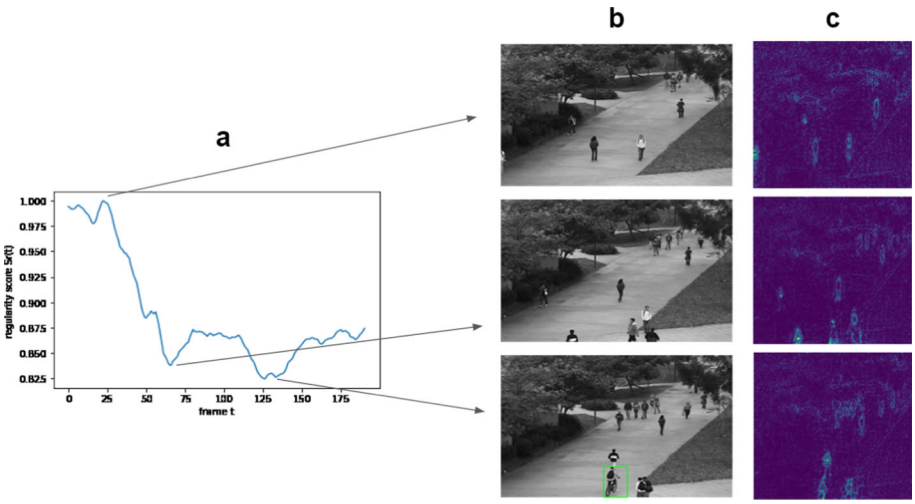

Fig. 22 Results of Conv-Lstm-AE on Ped1 (clip 016). (a) Regularity score of the clip. (b) Input frames. (c) Reconstruction error map for the input frames

object, rather than other factors (such as BG or the number of objects, etc.). In order to validate this idea, we separately trained and evaluated three autoencoders with, respectively, cropped images of objects and their class-level features, which were extracted by a pre-trained CNN. For this experiment, two groups of images were prepared and named 'Normals' and 'Abnormals'. The Normals group contains all of the cropped objects of the same group (here,

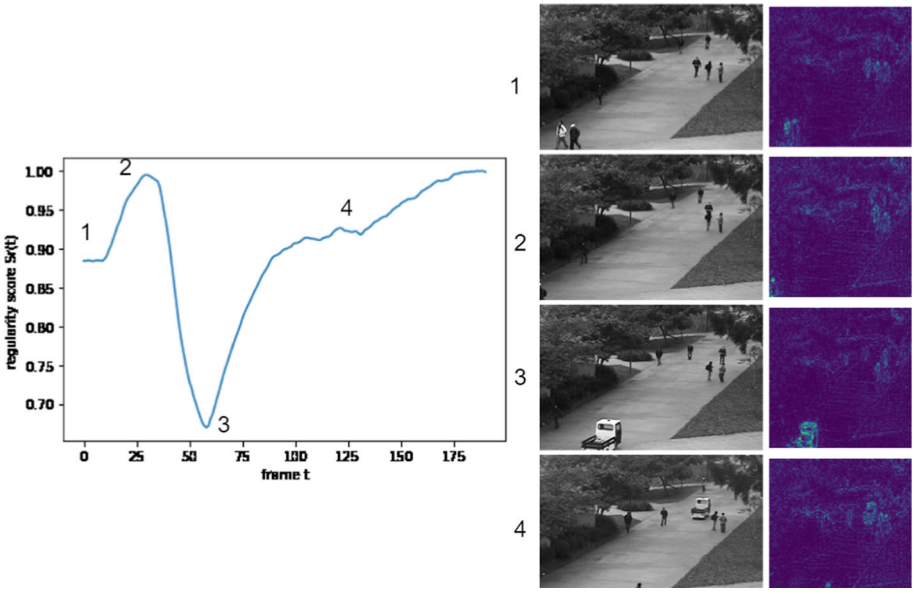

Fig. 23 Results of Conv-Lstm-AE on Ped1 (clip 020). (a) Regularity score of the clip. (b) Input frames. (c) Reconstruction error map for the input frames

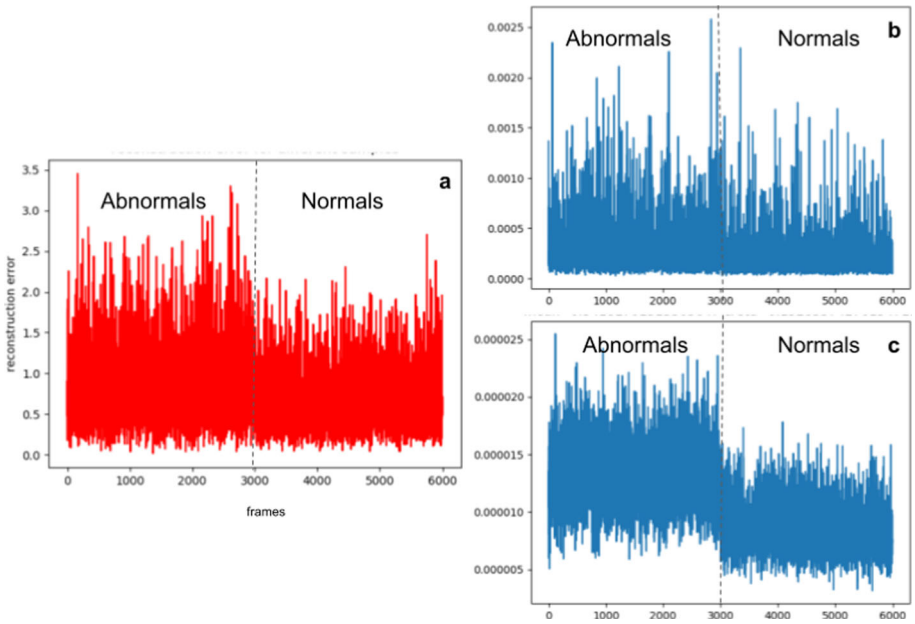

Fig. 24 Reconstruction error for Normal and Abnormal samples. (a) Reconstruction error for images. (b) Reconstruction error for low-level features. (c) Reconstruction error for class-level features

images of people) and the Abnormals group contains cropped images of different types of objects (such as vehicles, bicycles, bikes, etc.). The Normal and abnormal groups present normal and abnormal objects, respectively. Then, using a pre-trained VGG19, the class-level (i.e., features of the last layer), and color-level features (i.e., features of the first layer) were extracted for each group. Each network was trained and evaluated separately with images of normal objects, and their low-level and class-level features. The evaluation is carried out on

Table 4 Comparison of different DL-based AD strategies from different viewpoints

Strategy used for AD	Reconstruction	Prediction	Segmentation	Object-Centric	Memorization
Object class awareness	No	No	Yes	No	No
Generalization of model to anomalies	possible	possible	theoretically NO	possible	No
Extracted features	ST	ST	ST	ST	ST
Used DNNs	AE, Conv-AE	Conv-LSTM-AE, GANs,Unet	GANs, Unet	AE, Conv-AE	AE, Conv-AE
Aware of environment and contextual information (location, time)	Implicitly	Implicitly	Implicitly	No	Implicitly
Typical examples	[24, 33, 44]	[51, 64, 123]	[11, 12, 56]	[16, 48, 50, 109, 126]	[43, 85]

ST stands for Spatial and spatiotemporal respectively

both normal and abnormal groups. The reconstruction error for each group is calculated on the test subset, to evaluate their performance in discriminating between normal and anomalies. The results are shown in Fig. 24 which confirms the aforementioned idea.

Results in Fig. 24 show that the networks trained on images of the normal objects (Fig.e 24-a), or their low-level features (Fig. 24-b) cannot effectively discriminate between normal and abnormal objects. Although frame reconstruction shows better results (in discriminating between normal and anomalies) compared to the low-level feature reconstruction method, it is not effective for anomaly detection. We believe that in frame reconstruction, the network implicitly considers some other features from the image, in addition to pixel intensity or color. However, the results of Fig. 24-c show that training the network on class-level features (directly providing high-level features) helps the network effectively discriminate between normal and abnormal objects and improves performance. As it can be concluded from this experiment, object-centric approaches do not effectively consider the class of the objects for discriminating between normal and abnormal objects.

Before summarizing the results, it is worth mentioning that these experiments do not target all proposed methods in the mentioned categories. However, they show that all these methods should consider the mentioned points (also summarized below) to reduce the false positives and false negatives. In summary, our experiments highlight these points:

1) Methods that focus on low-level features (in their loss or target functions) do not effectively discriminate between normal and abnormal frames which are defined based on the class of the object. These methods are also vulnerable to illumination changes.

Table 5 Advantages and challenges of different DL-based semi-supervised VAD methods

AD strategy	Strong points	shortcomings
Reconstruction based [24, 33, 44, 70, 84, 118, 137]	*Implicitly aware of the environment and context (e.g., location, background, etc.).	*Unaware of the object class.
	*Considers low-level featurs (such as color and texture) for anomaly detection.	*The model would be generalized to abnormals.
Prediction based [22, 55, 64, 94, 100, 123, 131]	*Implicitly aware of the environment.	*Unaware of the object class.
	*nherently models motion patterns.	*The model would be generalized to abnormals.
Object centric methods [16, 31, 48, 50, 109, 126]	*Focuses on the objects.	*The performance is dependent on the object detection step.
	*anomaly score is not affected by the number of objects.	*Unaware of the environment.
Segmentation based [11, 12]	*Is aware of the class of the object.	*The performance is dependent on the semantic segmentation step.
	*Model is not distacted by background complexity.	*It has a higher computational load.
Memorization based [43, 65, 85, 107, 122]	Model is not generalizable to abnormals.	*Produced anomaly scores would be equal for different frames if their latent spaces are close to each other.

2) Methods that consider and analyze the frames entirely, instead of analyzing each object individually, will fail when the number of foreground objects is considerably variable in different frames.
3) Object-Centric methods are not affected by the number of foreground objects. However, they do not consider the environment information (BG information, location, etc.). Moreover, although they focus on objects (rather than redundant information such as BG), they are not aware of the class of the object if they reconstruct the cropped image in order to learn appearance patterns.
4) It is probable that in single-path methods, which analyze appearance and motion simultaneously, the motion information may be dominated by the appearance. Two branch approaches, analyze the appearance and motion separately; hence their effect can be applied separately to the task.
5) The effect of the camera distance (or object size) should be considered in the methods. Objects closer to the camera usually have more effect on the frame anomaly score.

Table 6 Some of the state-of-the-art DL-based semi-supervised VAD methods

State of the art methods	Strategy for AD	DNN used in the method	Special points
Hasan et al. [44]	Reconstruction	Conv-AE	*The proposed model (Conv-AE) does not consider temporal patterns effectively. *Not aware of the class of the objects. *Has difficulty detecting anomalies in small regions.
Chong et al. [24]	Reconstruction	Conv-LSTM-AE	*Considers the evolution of frames. *Not aware of the class of the objects.
Akcay et al. [6]	Reconstruction	Conditional GAN	*In addition to minimizing the distance between input and reconstructed images, the distance between latent spaces is also minimized.
Ravanbakhsh [93]	Reconstruction	GAN	*Benefits from generating optical flow images from raw-pixel frames and vice versa for AD.
Liu et al. [64]	Prediction	GAN (Unet for generator)	*Prediction considers the motion by modeling the evolution of frames. *Not aware of the class of the objects.
Ionescu et al. [48]	Object centric	AE	*Focuses on the objects but does not consider the environment. *The performance is dependent on the performance of the object detector.
Gong et al. [43]	Memorization	AE	*Benefits from a memory module and hence AE, here, is not generalizable to anomalies.
Park et al. [85]	Memorization	AE	*Considers the diversity of normal patterns explicitly. *AE, here, is not generalizable to anomalies.
Mohammad et al. [11]	Segmentation	Unet	*Considers the class of objects for VAD. *Motion is modeled by predicting optical flow magnitude.

5 Conclusion

In this study, recent DL-based semi-supervised video anomaly detection methods are reviewed. As DNNs are the main tool for different parts of the task (e.g., feature extraction, decision making), the study began with a study of DNNs. Different DNNs are reviewed and analyzed from different points of view, such as spatiotemporal feature extraction, pattern learning, and compatibility with different data types. Moreover, their applicability for different parts of anomaly detection methods is stated, providing some points regarding their special attributes and challenges. Hence, researchers can choose the most suitable DNN for different parts of their anomaly detection method, based on their approaches. In the final section, recent anomaly detection methods are critically reviewed. First, the methods are categorized based on the spatiotemporal feature extraction process and then the study analyzed them based on the strategy they commonly used for anomaly detection. Moreover, almost all of the recent approaches and state-of-the-art methods in the field are covered in this review, thereby providing a global but comprehensive look at the field for researchers, by describing essentials, positives points, shortcomings, and challenges of each categorization and approach, which can be the subject of future work. Tables 4 and 5 summarize different reviewed anomaly detection strategies and Table 6 presents the state-of-the-art research in this field.

Effective detection of video anomalies (similar to other applications requiring a good understanding of videos) requires joint consideration of different requirements, such as extracting effective appearance features, capturing motion and extracting effective temporal features, and separately capturing and analyzing different moving objects in the scene (for most applications), considering the context and the environment information, etc. However, each proposed method has addressed only one or a few of the mentioned requirements and almost all methods are unable to effectively and jointly consi der all of the aspects. This issue should be addressed in future work by properly combining existing methods, considering their capabilities and the capabilities of different DNNs. Initial efforts in this direction are made in our recent research in video anomaly detection [11, 12] and it is our hope many other researchers will join us on this endeavor.

Acknowledgements This work was supported by a Discovery Grant (RGPIN-2016-05876) from the National Sciences and Engineering Research Council of Canada (NSERC). Additionally, we would like to thank Annette Schwerdtfeger for proofreading the paper.

Author Contributions The idea for this article was proposed by Mohammad Baradaran (the principal author), who also performed the literature search and data analysis, and drafted the work. This work was supervised by Robert Bergevin who also critically revised the article.

Data Availability The datasets analyzed during the current study are publicly available. All needed references are provided in the document.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. (2015) SIIM-ACR Pneumothorax Segmentation. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/discussion/107981

2. (2021) Unusual crowd activity dataset of University of Minnesota, available at: http://mha.cs.umn.edu/movies/crowdactivity-all.avi
3. Abati D, Porrello A, Calderara S, Cucchiara R (2019) Latent space autoregression for novelty detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, pp 481–490
4. Aburakhia S, Tayeh T, Myers R, Shami A (2020) A transfer learning framework for anomaly detection using model of normality. The 11th Annual IEEE Information Technology, Electronics and Mobile Communication Conference "IEEE IEMCON", Vancouver, Canada
5. Adam A, Rivlin E, Shimshoni I, Reinitz D (2008) Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(03):555–560
6. Akcay S, Atapour-Abarghouei A, Breckon TP (2018) GANomaly: Semi-supervised anomaly detection via adversarial training. In: Jawahar C, Li H, Mori G, Schindler K (eds) Computer Vision - ACCV 2018, vol 11363. Lecture Notes in Computer Science. Springer, Cham
7. Akçay S, Atapour-Abarghouei A, Breckon TP (2019) Skip-GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, pp 1–8. https://doi.org/10.1109/IJCNN.2019.8851808
8. Alaslani MG, Elrefaei LA (2018) Convolutional neural network based feature extraction for IRIS recognition. International Journal of Computer Science and Information Technology (IJCSIT) 10(2)
9. Alkhayrat M, Aljnidi M, Aljoumaa KA (2020) Comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. J Big Data 7(9). https://doi.org/10.1186/s40537-020-0286-0
10. Arif S, Wang J, Hassan TU, Fei Z (2019) 3D-CNN-Based fused feature maps with LSTM applied to action recognition. Future Internet. https://doi.org/10.3390/fi11020042
11. Baradaran M, Bergevin R (2022) Object class aware video anomaly detection through image translation. 2022 19th Conference on Robots and Vision (CRV), pp 90–97. https://doi.org/10.1109/CRV55824.2022.00020
12. Baradaran M, Bergevin R (2023) Multi-task learning based video anomaly detection with attention. CVPRW-VAND
13. Baur C, Wiestler B, Albarqouni S, Navab N (2018) Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T (eds) Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. BrainLes, lecture notes in computer science, vol 11383. Springer
14. Bergmann P, Fauser M, Sattlegger D, Steger C (2020) Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. CVPR
15. Beula Rani BJ, Sumathi L, E M (2020) Survey on applying GAN for anomaly detection. 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp 1–5. https://doi.org/10.1109/ICCCI48352.2020.9104046
16. Biradar KM, Gupta A, Mandal M, Vipparthi SK (2019) Challenges in time-stamp aware anomaly detection in traffic videos. IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)
17. Bulusu S, Kailkhura B, Li B, Varshney PK, Song D (2020) Anomalous example detection in deep learning: A survey. IEEE Access 8:132330–132347. https://doi.org/10.1109/ACCESS.2020.3010274
18. Carreira J, Zisserman A, Vadis Q (2017) Action recognition: a new model and the kinetics dataset. Proc IEEE Conf Comput Vis Pattern Recognit 6299–6308
19. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: A survey. ACM Comput Surv 41(3):58. Article 15. https://doi.org/10.1145/1541880.1541882
20. Chang X, Zhang Y, Xue D, Chen D (2022) Multi-task learning for video anomaly detection. J Vis Commun Image Represent 87. https://doi.org/10.1016/j.jvcir.2022.103547
21. Chen Y, Kalantidis Y, Li J, Yan S, Feng J (2018) Multi-fiber networks for video recognition. Proceedings of the European Conference on Computer Vision (ECCV), pp 352–367
22. Chen D, Wang P, Yue L, Zhang Y, Jia T (2020) Anomaly detection in surveillance video based on bidirectional prediction. Image Vis Comput V 98
23. Chen C, Yuan W, Xie Y, Qu Y, Tao Y, Song H, Ma L (2020) Novelty detection via non-adversarial generative network. arXiv:2002.00522
24. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. In: Cong F, Leung A, Wei Q (eds) Advances in Neural Networks. ISNN 2017. Lecture Notes in Computer Science, vol 10262. Springer
25. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS 2014 Workshop on Deep Learning

26. Cinelli LP, Thomaz LA, da Silva AF, da Silva EAB, Netto SL (2017) Foreground segmentation for anomaly detection in surveillance videos using deep residual networks. São Pedro-Brazil, XXXV Simpósio Brasileiro de Telecomunicações e Processamento de SinaisAt
27. Dargan S, Munish K, Ayyagari MR, Gulshan k (2020) A survey of deep learning and its applications: A new paradigm to machine learning. Archives of Computational Methods in Engineering 1–22
28. Di MF, Galeone P, Simoni MD, Ghelfi E (2019) A survey on GANs for anomaly detection. arXiv:1906.11632
29. Djuris J, Ibric S, Djuric Z (2013) Neural computing in pharmaceutical products and process development. Computer-Aided Applications in Pharmaceutical Technology 91–175
30. Donahue J, Krhenbhl P, Darrell T (2017) Adversarial feature learning. International Conference on Learning Representations (ICLR)
31. Doshi K, Yilmaz Y (2020) Any-shot sequential anomaly detection in surveillance videos. CVPR
32. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. 9th International conference on learning representations, ICLR
33. Duman E, Erdem OA (2019) Anomaly detection in videos using optical flow and convolutional autoencoder. IEEE Access 7:183914–183923. https://doi.org/10.1109/ACCESS.2019.2960654
34. Dzmitry B, Cho, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. ICLR2015
35. Fang Z, Zhou JT, Xiao Y, Li Y, Yang F (2021) Multi-encoder towards effective anomaly detection in videos. In: IEEE Transactions on Multimedia. https://doi.org/10.1109/TMM.2020.3037538
36. Fan Y, Lu X, Li D, Liu Y (2016) Video-based emotion recognition using CNN-RNN and C3D Hybrid Networks. ICMI '16, November 12-16, Tokyo, Japan
37. Feichtenhofer C, Pinz A, Wildes RP (2020) Deep insights into convolutional networks for video recognition. Int J Comput Vis 128:420–437
38. Feichtenhofer C, Pinz A, Wildes RP (2016) Spatiotemporal residual networks for video action recognition. NIPS
39. Ganokratanaa T, Aramvith S, Sebe N (2020) Unsupervised anomaly detection and localization based on deep spatiotemporal translation network. IEEE Access 8:50312–50329. https://doi.org/10.1109/ACCESS.2020.2979869
40. Georgescu MI, Bărbălău A, Ionescu RT, Shahbaz Khan F, Popescu M, Shah M (2021) Anomaly detection in video via self-supervised and multi-task learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12737–12747. https://doi.org/10.1109/CVPR46437.2021.01255
41. Gherbi E, Hanczar B, Janodet J, Klaudel W (2019) An encoding adversarial network for anomaly detection. Proceedings of The Eleventh Asian Conference on Machine Learning, PMLR 101:188–203
42. Gondara L (2016) Medical image denoising using convolutional denoising autoencoders. In: Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW), pp 241–246
43. Gong D, Liu L, Le V, Saha B, Mansour MR, Venkatesh S, Hengel Avd (2019) Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. ICCV
44. Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp 733–742
45. Hinami R, Mei T, Satoh S (2017) Joint detection and recounting of abnormal events by learning deep generic knowledge. 2017 IEEE International Conference on Computer Vision (ICCV). Venice, pp 3639–3647. https://doi.org/10.1109/ICCV.2017.391
46. Ho K, Keuper J, Keuper Mt (2020) Unsupervised multiple person tracking using autoencoder-based lifted multicuts. arXiv:2002.01192
47. Houssam Z, Chuan F, Bruno L, Gaurav M, Vijay C (2018) Efficient GAN-based anomaly detection. arXiv:1802.06222
48. Ionescu RT, Khan FS, Georgescu M, Shao L (2019) Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, pp 7834–7843
49. Ionescu RT, Smeureanu S, Alexe B, Popescu M (2017) Unmasking the abnormal events in video. 2017 IEEE International conference on computer vision (ICCV), Venice, pp 2914–2922. https://doi.org/10.1109/ICCV.2017.315
50. Jianfei Z, Yi Z, Pan S, Zhao Y, Zhao Z, Su F, Zhuang B (2019) Unsupervised traffic anomaly detection using trajectories. CVPR Workshops
51. Jones MJ, Ramachandra B (2020) Street Scene: A new dataset and evaluation protocol for video anomaly detection. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

52. Kanojia G, Kumawat S, Raman S (2019) Exploring temporal differences in 3D convolutional neural networks. In: Babu RV, Prasanna M, Namboodiri VP (eds) Computer vision, pattern recognition, image processing, and graphics. NCVPRIPG 2019. Communications in computer and information science, vol 1249. Springer, Singapore. https://doi.org/10.1007/978-981-15-8697-2-10
53. Khan A, Sohail A, Zahoora U, Qureshi AS (2019) A survey of the recent architectures of deep convolutional neural networks. arXiv:1901.06032
54. Kimura M, Yanagihara T (2018) Anomaly detection using GANs for visual inspection in noisy training data. In: Carneiro G, You S (eds) Computer Vision - ACCV 2018 Workshops, vol 11367. Lecture Notes in Computer Science. Springer, Cham
55. Kiran BR, Thomas DM, Parakkal R (2018) An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. J Imaging 4(36)
56. Krzysztof L, Nakka KK, Fua P, Salzmann M (2019) Detecting the unexpected via image resynthesis. IEEE/CVF International Conference on Computer Vision (ICCV), pp 2152–2161
57. Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogianni I (2018) Artificial neural networks and machine learning. ICANN 2018: 27th International conference on artificial neural networks, Rhodes, Greece
58. Le XH, Ho HV, Lee G, Jung S (2019) Application of Long Short-Term Memory (LSTM) neural network for flood forecasting. MDPI, water
59. Lee JY, Nam WJ Lee SW (2022) Multi-contextual predictions with vision transformer for video anomaly detection. arXiv:2206.08568
60. Lee H, Grosse R, Ranganath R, Ng AY (2011) Unsupervised learning of hierarchical representations with convolutional deep belief networks. Commun ACM 54(10):95–103. https://doi.org/10.1145/2001269.2001295
61. Leng L, Zhang J (2013) PalmHash Code vs. PalmPhasor Code. Neurocomputing 108:1–2. https://doi.org/10.1016/j.neucom.2012.08.028
62. Leng L, Li M, Kim C et al (2017) Dual-source discrimination power analysis for multi-instance contactless palmprint recognition. Multimed Tools Appl 76:333–354. https://doi.org/10.1007/s11042-015-3058-7
63. Li Z, Li Y, Gao Z (2020) Spatiotemporal representation learning for video anomaly detection. IEEE Access 8:25531–25542
64. Liu W, Luo W, Lian D, Gao S (2018) Future frame prediction for anomaly detection - a new baseline. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, pp 6536–6545
65. Liu Z, Nie Y, Long C, Zhang Q, Li G (2021) 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE/CVF International Conference on Computer Vision (ICCV), pp 13568–13577. https://doi.org/10.1109/ICCV48922.2021.01333
66. Lu X, Wang W, Shen J, Crandall D, Luo J (2022) Zero-shot video object segmentation with co-attention siamese networks. IEEE Trans Pattern Anal Mach Intell 44(4):2228–2242. https://doi.org/10.1109/TPAMI.2020.3040258
67. Lu Y, Kumar KM, Nabavi SS, Wang Y (2019) Future frame prediction using convolutional vrnn for anomaly detection. 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp 1–8. https://doi.org/10.1109/AVSS.2019.8909850
68. Luo W, Liu W, Lian D, Gao S (2021) Future frame prediction network for video anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). https://doi.org/10.1109/TPAMI.2021.3129349
69. Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked RNN framework. 2017 IEEE International Conference on Computer Vision (ICCV). Venice, pp 341–349. https://doi.org/10.1109/ICCV.2017.45
70. Luo W, Liu W, Gao S (2017) Remembering history with convolutional LSTM for anomaly detection. IEEE International conference on multimedia and expo, ICME 2017. Hong Kong, China, July 10-14, pp 439–444
71. Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1412–1421
72. Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. ICCV
73. Lu C, Shi J, Jia J (2013) Abnormal Event Detection at 150 FPS in Matlab. Int Conf Comput Vis (ICCV)
74. Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F (2019) See more, know more: Unsupervised video object segmentation with co-attention siamese networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp 3618–3627. https://doi.org/10.1109/CVPR.2019.00374

75. Mahadevan V, Li W, Bhalodia V, Vasconcelos N (2010) Anomaly detection in crowded scenes. Proc IEEE Conf Comput Vis Pattern Recognit(CVPR) 1975–1981
76. Manassés R, André L, Heitor L (2018) A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recognit Lett 105:13–22
77. Medel JR, Savakis A (2016) Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv:1612.00390
78. Medel JR, Savakis A (2016) Anomaly detection in video using predictive convolutional long short-term memory networks. arXiv:1612.00390
79. Métais E, Meziane F, Vadera S, Sugumaran V, Saraee M (2019) Natural language processing and information systems.(book): 24th International conference on applications of natural language to information systems, NLDB 2019, Salford, UK
80. Minderer M, Sun C, Villegas R, Cole F, Murphy K, Lee H (2019) Unsupervised learning of object structure and dynamics from videos. 33rd Conference on Neural Information Processing Systems (NeurIPS)
81. Morais R, Le V, Tran T, Saha B, Mansour M, Venkatesh S (2019) Learning regularity in skeleton trajectories for anomaly detection in videos. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, pp 11988–11996
82. Narasimhan MG, S SK (2018) Dynamic video anomaly detection and localization using sparse denoising autoencoders. Multimed Tools Appl 77:13173–13195
83. Nazare TS, deMello RF, Ponti MA (2018) Are pre-trained CNNs good feature extractors for anomaly detection in surveillance videos? arXiv:1811.08495
84. Nguyen TN, Meunier J (2019) Anomaly detection in video sequence with appearance-motion correspondence. ICCV
85. Park H, Noh J, Ham B (2020) Learning memory-guided normality for anomaly detection. CVPR
86. Pascanu R, Mikolov T, Bengio Y (2013) On the difficulty of training recurrent neural networks. Proceedings of the 30th International Conference on Machine Learning, PMLR 28(3):1310–1318
87. Pathak D, Girshick R, Dollár P, Darrell T, Hariharan B (2017) Learning features by watching objects move. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp 6024–6033. https://doi.org/10.1109/CVPR.2017.638
88. Pihlgren GG, Sandin F, Liwicki M (2020) Improving image autoencoder embeddings with perceptual loss. IJCNN/WCCI
89. Pinggera P, Ramos S, Gehrig S, Franke U, Rother C, Mester R (2016) Lost and found: Detecting small road hazards for self-driving vehicles. Proceedings of IROS 2016, Daejeon, Korea
90. Raghavendra C, Sanjay C (2019) Deep learning for anomaly detection: A survey. arXiv:1901.03407
91. Ramachandra B, Jones MJ, Vatsavai RR (2022) A survey of single-scene video anomaly detection. IEEE Trans Pattern Anal Mach Intell 44:2293–2312
92. Ramaswamy A, Seemakurthy K, Gubbi J, Purushothaman B (2020) Spatio-temporal action detection and localization using a hierarchical LSTM. CVPR workshop
93. Ravanbakhsh M, Nabi M, Sangineto E, Marcenaro L, Regazzoni C, Sebe N (2017) Abnormal event detection in videos using generative adversarial nets. IEEE International Conference on Image Processing (ICIP), Beijing, pp 1577–1581
94. Reiter W (2020) Video anomaly detection in post-procedural use of laparoscopic videos. In: Tolxdorff T, Deserno T, Handels H, Maier A, Maier-Hein K, Palm C (eds) Bildverarbeitung für die Medizin 2020. Informatik aktuell. Springer Vieweg, Wiesbaden
95. Ren J, Xia F, Liu Y, Lee I (2021) Deep video anomaly detection: Opportunities and challenges. In: 2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand, pp 959–966. https://doi.org/10.1109/ICDMW53433.2021.00125
96. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y (2011) Contractive auto-encoders: Explicit invariance during feature extraction. Proceedings of the 28th international conference on machine learning (ICML-11), pp 833–840
97. Ristea NC et al. (2022) Self-supervised predictive convolutional attentive block for anomaly detection. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13566–13576. https://doi.org/10.1109/CVPR52688.2022.01321.
98. Roka S, Diwakar M, Singh P, Singh P (2023) Anomaly behavior detection analysis in video surveillance: a critical review. J Electron Imaging 32(4):042106. https://doi.org/10.1117/1.JEI.32.4.042106
99. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. Int Conf Med Image Comput Comput Assist Interv 234–241. Springer
100. Roy PR, Bilodeau GA, Seoud L (2020) Local anomaly detection in videos using object-centric adversarial learning. The First International Workshop on Deep Learning for Human-Centric Activity Understanding (ICPR2020 workshop)

101. Sabokrou M (2018) AVID: Adversarial Visual Irregularity Detection. ACCV, Lecture notes in computer science, vol 11366. Springer
102. Sabokrou M, Fayyaz M, Fathy M, Klette R (2017) Deepcascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Trans Image Process 26(4):1992–2004
103. Sabokrou M, Khalooei M, Fathy M, Adeli E (2018) Adversarially learned one-class classifier for novelty detection. In: Proc CVPR
104. Sellat H (2019) Anomaly detection in videos using LSTM convolutional autoencoder. Available at: https://towardsdatascience.com/prototyping-an-anomaly-detection-system-for-videos-step-by-step-using-lstm-convolutional-4e06b7dcdd29
105. Sengupta S, Basak S, Saikia P, Paul S, Tsalavoutis V, Atiah FD, Ravi V, Peters RA (2020) A review of deep learning with special emphasis on architectures, applications and recent trends. Knowl-Based Syst 194
106. Shafkat I (2019) Intuitively understanding variational autoencoders. www.towardsdatascience.com
107. Shen G, Ouyang Y, Sanchez V (2022) Video anomaly detection via prediction network with enhanced spatio-temporal memory exchange. 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 3728–3732. https://doi.org/10.1109/ICASSP43922.2022.9747376
108. Shibin P, Josh H, Christopher B, Scott S, Michael R (2016). Evaluation schemes for video and image anomaly detection algorithms. Proceedings of the SPIE, Vol. 9844
109. Shine L, Edison A, Jiji CV (2019) A comparative study of faster R-CNN models for anomaly detection in 2019 AI City Challenge. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp 306–314
110. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), vol 1, pp 568–576
111. Singh H (2019) Anomalous motion detection of vehicles on highways using deep learning. University of Nevada, Reno, Thesis
112. Smys S, Tavares JMRS, Balas VE, Iliyasu AM (2019) Computational vision and bio-inspired computing. (book) Springer International Publishing, ICCVBIC 2019, series vol. 1108
113. Srivastava N, Mansimov E, Salakhutdinov R (2015) Unsupervised learning of video representations using LSTMs. In: ICML
114. Sultani W, Chen C, Shah M (2018) Real-World anomaly detection in surveillance videos. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp 6479–6488. https://doi.org/10.1109/CVPR.2018.00678
115. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. IEEE/CVF Conference on computer vision and pattern recognition, Salt Lake City, pp 6479–6488
116. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. 2015 IEEE International Conference on Computer Vision (ICCV), pp 4489–449
117. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT pp 6450–6459. https://doi.org/10.1109/CVPR.2018.00675
118. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. 2018 IEEE/CVF Conference on computer vision and pattern recognition. Salt Lake City, UT,pp 6450–6459. https://doi.org/10.1109/CVPR.2018.00675
119. Tuan TX, Phuong TM (2017) 3D Convolutional networks for session-based recommendation with content features. The Eleventh ACM Conference
120. Tuan HV, Sebastien A, Jacues B, Abdelmalik TA (2020) Anomaly detection in surveillance videos by future appearance-motion prediction. Proc 15th Int Jt Conf Comput Vis. Imaging Comput Graph Theory Appl 5:484–490. https://doi.org/10.5220/0009146704840490
121. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Aidan AN, K U, Polosukhin I (2017) Attention is All you Need. Neurips 30
122. Wang B, Yang C (2022) Video anomaly detection based on convolutional recurrent autoencoder. Sensors 22. https://doi.org/10.3390/s22124647
123. Wang S, Cao J, Yu P (2022) Deep learning for spatio-temporal data mining: A survey. IEEE Trans Knowl Data Eng 34(08):3681–3700. https://doi.org/10.1109/TKDE.2020.3025580
124. Wang Y, Liao W, Chang Y (2018) Gated recurrent unit network-based short-term photovoltaic forecasting. Energies 2018 11(8)
125. Wang L, Zhou F, Li Z, Zuo W, Tan H (2018) Abnormal event detection in videos using hybrid spatio-temporal autoencoder. 2018 25th IEEE International Conference on Image Processing (ICIP). Athens, pp 2276–2280. https://doi.org/10.1109/ICIP.2018.8451070

126. Wei J, Zhao J, Zhao Y, Zhao Z (2018) Unsupervised anomaly detection for traffic surveillance based on background modeling. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Salt Lake City, pp 129–1297
127. Xu M, Yu X, Chen D, Wu C, Jiang Y (2019) An efficient anomaly detection system for crowded scenes using variational autoencoders. Appl Sci 9(16):3337
128. Xu D, Ricci E, Yan Y, Song J, Sebe N (2015) Learning deep representations of appearance and motion for anomalous event detection. Proceedings of the British Machine Vision Conference (BMVC)
129. Yadav RK, Kumar R (2022) A Survey on video anomaly detection. 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, pp 1–5. https://doi.org/10.1109/DELCON54057.2022.9753580
130. Ye W, Cheng J, Yang F, Xu Y (2019) Two-stream convolutional network for improving activity recognition using convolutional long short-term memory networks. IEEE Access 7:67772–67780. https://doi.org/10.1109/ACCESS.2019.2918808
131. Ye M, Peng X, Gan W, Wu W, Qiao Y (2019) AnoPCN: Video anomaly detection via deep predictive coding network. In: Proceedings of the 27th ACM International Conference on Multimedia (MM '19). Association for Computing Machinery, New York, NY, USA, pp 1805–1813. https://doi.org/10.1145/3343031.3350899
132. Yiru Z, Bing D, Chen S, Yao L, Hongtao L, Xian-Sheng H (2017) Spatio-temporal autoencoder for video anomaly detection. ACM Multimedia Conference
133. Yuan H, Cai Z, Zhou H, Wang Y, Chen X (2021) TransAnomaly: Video anomaly detection using video vision transformer. IEEE Access 9:123977–123986. https://doi.org/10.1109/ACCESS.2021.3109102
134. Zaheer MZ, Mahmood A, Khan MH, Segu M, Yu F, Lee S-I (2022) Generative cooperative learning for unsupervised video anomaly detection. IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR) 2022:14724–14734. https://doi.org/10.1109/CVPR52688.2022.01433
135. Zenati H, Romain M, Foo C, Lecouat B, Chandrasekhar V (2018) Adversarially learned anomaly detection. 2018 IEEE International Conference on Data Mining (ICDM), Singapore, pp 727–736. https://doi.org/10.1109/ICDM.2018.00088
136. Zhang Y, Nie X, He R, Chen M, Yin Y (2020) Normality learning in multispace for video anomaly detection. IEEE Transaction on Circuits and Systems for Video Technology. https://doi.org/10.1109/TCSVT.2020.3039798
137. Zhang C, Li S, Zhang H, Chen Y (2020) VELC: A new variational autoencoder based model for time series anomaly detection. arXiv:1907.01702

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.