Anomaly Analysis in Images and Videos: A Comprehensive Review

TUNG MINH TRAN, TU N. VU, and NGUYEN D. VO, University of Information Technology, Ho Chi Minh City, Vietnam and Vietnam National University, Ho Chi Minh City, Vietnam
TAM V. NGUYEN, University of Dayton, Dayton, Ohio, USA
KHANG NGUYEN, University of Information Technology, Ho Chi Minh City, Vietnam and Vietnam National University, Ho Chi Minh City, Vietnam

Anomaly analysis is an important component of any surveillance system. In recent years, it has drawn the attention of the computer vision and machine learning communities. In this article, our overarching goal is thus to provide a coherent and systematic review of state-of-the-art techniques and a comprehensive review of the research works in anomaly analysis. We will provide a broad vision of computational models, datasets, metrics, extensive experiments, and what anomaly analysis can do in images and videos. Intensively covering nearly 200 publications, we review (i) anomaly related surveys, (ii) taxonomy for anomaly problems, (iii) the computational models, (iv) the benchmark datasets for studying abnormalities in images and videos, and (v) the performance of state-of-the-art methods in this research problem. In addition, we provide insightful discussions and pave the way for future work.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Anomalies, anomaly analysis, anomaly detection, anomaly prediction, deep learning

ACM Reference format:
Tung Minh Tran, Tu N. Vu, Nguyen D. Vo, Tam V. Nguyen, and Khang Nguyen. 2022. Anomaly Analysis in Images and Videos: A Comprehensive Review. *ACM Comput. Surv.* 55, 7, Article 148 (December 2022), 37 pages.
https://doi.org/10.1145/3544014

1 INTRODUCTION

1.1 Study Background

Nowadays, to ensure human safety and prevent anomalous actions or events, such as criminal behaviors, traffic accidents, burglary, and fighting, smart surveillance systems [160] are being increasingly installed to raise alarms in potentially dangerous situations. Anomaly detection and

This project is partially funded by National Science Foundation (NSF) under Grant No. 2025234 and Vietnam National University Ho Chi Minh City (VNU-HCM) under Grant No. DS2021-26-01.
Authors' addresses: T. M. Tran, T. N. Vu, N. D. Vo, and K. Nguyen, Quarter 6, Linh Trung Ward, Thu Duc District, University of Information Technology, Ho Chi Minh City, Vietnam and Vietnam National University, Ho Chi Minh City, Vietnam; emails: tungtm.ncs@grad.uit.edu.vn, 18520184@gm.uit.edu.vn, nguyenvd@uit.edu.vn, khangnttm@uit.edu.vn; T. V. Nguyen, 300 College Park, University of Dayton, Dayton, 45469, OH, USA; email: tamnguyen@udayton.edu.

0360-0300/2022/12-ART148 $15.00
https://doi.org/10.1145/3544014

localization are becoming more significant in the manufacturing industry and medicine for detecting product faults and diagnosing diseases. Anomaly analysis in images and videos is crucially challenging due to their high-dimensional structure of the images, combined with the non-local temporal variations across frames. In addition, there are domain challenges in the real world such as various environmental conditions (illumination variations, shadow effects of the objects, object occlusions, and cluttered backgrounds), crowd density, the complex nature of human behaviors, recording camera setting with uncontrolled, and difficulty in accessing good computational infrastructure. In particular, the anomaly datasets are of extreme imbalance make anomaly analysis in images and videos one of the daunting tasks in the computer vision and machine learning fields.

Furthermore, real-world anomalous actions or events are complicated and diverse, since the environment captured by surveillance cameras can change drastically over time. Therefore, it is challenging to list all the possible abnormal or suspicious activities in the real world. Abnormal activities can be identified as illegal activities from normal ones. Hence, it attracts a significant amount of works in the computer vision community with a lot of applications about anomaly problems, including anomaly detection, anomaly classification, anomaly prediction, anomaly localization from images and videos. Though there is a huge amount of works about unusual problems and some surveys on this topic [3, 17, 54, 90, 93, 125, 159]. Different from the previous works, which commonly concentrate on studying problems related to image or video anomalies, our study further focuses on investigating anomaly analysis tasks in both image and video domains, especially for methods dealing with human anomalous activity problems. To the best of our knowledge, there is no comprehensive review of this topic from images and videos yet.

1.2 Scope and Motivation

We have noted that the scope of the study should cover the nature of feature representation, the feasibility of various deep learning techniques and approaches, taxonomies of anomalous problems, benchmark datasets, suitability of the techniques in application contexts, action recognition and anomaly analysis outputs, and evaluation criteria in images and videos.

Our work is motivated by some aspects. First, we dwell on distinguishing between traditional methods based on manual features and those based on deep learning to highlight recent advances in deep learning techniques for action recognition and anomaly analysis in images and videos. Second, we identify challenges when dealing with anomaly problems such as anomaly detection, anomaly classification, anomaly prediction in images and videos, and the range of applications of these problems that existing surveys do not fully cover on this topic. Third, this article also compares the performances of various state-of-the-art techniques on the benchmark datasets to show the current state of research. Next, we carry out to experiment with three well-known methods on two anomalous datasets in images, namely, BTAD [98] and MTD [46], and perform an in-depth analysis of current approaches in real scenarios. Then, we also conduct to experiment with three prominent methods on three landmark datasets in videos, namely, Subway Entrance [1], UCF-Crime [150], and Street Scene [123], as well as outlining practical challenges in the development of methods of automatic visual analysis of abnormal activities in different contexts. Finally, we discuss the limitations of the state-of-the-art and present promising avenues for further research to understand anomalous human actions and behaviour.

The main contribution of the article is fivefold. First, we compile a comprehensive survey of anomaly analysis techniques from images and videos. Second, we clearly defined problem statements for anomaly analysis tasks in images and videos. Third, we contribute a coherent and systematic review of state-of-the-art techniques through pre-processing, feature extraction, and modelling in images and videos. Fourth, we review anomaly benchmark datasets that are available and commonly used in fields of computer vision, together with their strong and weak points.

Table 1. Summary of Previous Reviews in Images

#	Title	Venue	Description
1	Survey on blind image forgery detection [120]	IET Image Processing	Reviews various blind techniques to detect image forgeries focusing on three common forgery types, namely, copy-move, splicing and retouching.
2	A Survey on Image Forgery Detection Techniques [94]	Digital Image Processing	Reviews the classification of image forgery detection approaches into two main methods, namely, active methods and passive methods.
3	Image Forgery Detection: Survey and Future Directions [96]	Data, Engineering and applications	Reviews four main types of forgery detection techniques such as image splicing, copy-move, resampling, and retouching detection, and discusses to extend these techniques in videos.
4	Image Anomalies: A review and Synthesis of Detection Methods [31]	Journal of Mathematical Imaging and Vision	Presents a classification of the methods based on five groups (e.g., distance-based methods, reconstruction-based methods) emerging for the background model.
5	Deep Learning for Medical Anomaly Detection–A Survey [35]	ACM Computing Surveys	Reviews the types of data used in medicine, methods based on deep architectures (e.g., autoencoders, GAN, multi-task learning, long short-term memory) and limitations of existing deep medical anomaly detection techniques.

See Section 1.3 for more details.

These datasets designed for the evaluation of the various methodologies for action recognition problems as well as anomaly analysis in images and videos are illustrated in our article, comparing the current state-of-the-art methods. Last, we review the performance of the state-of-the-art methods, and extensive experiments, and discuss the research outlook.

1.3 Related Surveys

In the past few years, some survey papers relevant to research trends on the topic of action recognition and anomaly analysis dealing with images and videos have been published. In this section, we provide a broad variety of previous survey methods that have been proposed for the aforementioned problems. Additionally, a visual list of the recent referenced surveys for deep anomaly analysis in images and videos is compiled in Tables 1 and 2.

Regarding action recognition in images and videos, a survey of Reference [3] reviewed different types of human activities for activity recognition. Furthermore, hierarchical recognition methodologies for high-level activities (statistical approaches, syntactic approaches, and description-based approaches) and four common datasets were also discussed. Moreover, different approaches for human motion analysis using depth data (e.g., depth-based, skeleton-based activity recognition, facial feature detection, hand gesture recognition) and various transfer-based activity recognition techniques (sensor modality, labelled data, feature-representation transfer, and relational-knowledge transfer) were reviewed by References [22, 176], respectively. A survey by Reference [135] focused on the classification of human activity analysis based on feature extraction including initial extraction and action interpretation. In addition, the review also presented various techniques for human activity recognition and described six available datasets. Additionally, a similar taxonomy as in Reference [3] was applied for comparison with different methods. As noted in Reference [159], the aim of this survey provided video representation in terms of low-level features, mid-level features, and unsupervised features. In addition, the review also suggested human activity prediction techniques in both discriminative and generative models and described five benchmark datasets. Likewise, a study by Reference [42] provided various action recognition approaches in videos including handcrafted representation solutions (e.g., holistic representations,

Table 2. Summary of Previous Reviews in Videos

#	Title	Venue	Description
1	An Overview of Deep Learning-based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos [60]	Journal of Imaging	Surveys different methods for anomaly detection based on unsupervised and semi-supervised deep learning architectures, namely, representation learning for reconstruction, predictive modelling, and deep generative models.
2	A Survey of Single-Scene Video Anomaly Detection [125]	IEEE Transactions on Pattern Analysis and Machine Intelligence	Revisits different approaches for single-view video anomaly detection and provides a comprehensive comparison of these methods on eight benchmark video datasets.
3	Anomaly detection in road traffic using visual surveillance: A survey [136]	ACM Computing Surveys	Reviews deep learning-based methods for anomaly detection focusing on entities in road traffic scenarios (e.g., vehicles, pedestrian, environment).
4	A comprehensive review on deep learning-based methods for video anomaly detection [105]	Image and Vision Computing	Presents various methods for video anomaly detection using deep learning and discusses the performance of these methods both for quantitative and qualitative analyses.
5	Deep Learning for Anomaly Detection: A Review [113]	ACM Computing Surveys	Surveys various deep learning-based methods for anomaly detection focusing on three principled frameworks, namely, deep learning for generic feature extraction, learning representations of normality, and end-to-end anomaly score learning and categorizes these methods based on 11 different models.

See Section 1.3 for more details.

local representation-based approaches) and deep learning-based solutions (e.g., Spatio-temporal networks, multiple stream networks, deep generative models, temporal coherency networks). Furthermore, nine available datasets and the performance of pioneering methods were also discussed. Then, another survey by Reference [93] narrowly focused on different techniques for human activity prediction in videos and discussed the merits and demerits of these methods. A survey of different methods for RGB-D-based motion recognition using deep learning was introduced by Reference [166]. The survey mainly classified motion recognition methods depending on the different properties of the modalities into four groups including RGB-based, depth-based, skeleton-based, and RGB+D-based, and described 15 relevant datasets. A short survey by Reference [148] presented different methods for recognition and detection of human-human interactions based on hand-crafted features (e.g., local features approach, global features approach) and those based on deep learning (e.g., single frame network, motion-based, and stream networks, recurrent networks). In addition, this survey also summarized 11 publicly available datasets and discussed the limitations of the state-of-the-art. Finally, Reference [187] reviewed various prominent techniques for action recognition methods including action features representation, interaction recognition, and action detection methods, and summarized 12 popular datasets.

Regarding anomaly analysis in videos, the survey by Reference [18] presented types of anomalies including point anomaly, contextual anomaly, and collective/group anomalies depending on context and the environment. In addition, the review also provided various anomaly detection techniques (e.g., supervised anomaly detection, semi-supervised anomaly detection, unsupervised anomaly detection) and applications for this problem. The review paper in Reference [87] provided crowd scene analysis approaches, such as pixel-based analysis, texture-based analysis, deep learning-based analysis, and various anomaly detection methods in crowded scenes including vision-based method, physics-inspired method but this survey was not mentioned crowd

datasets. Later, Reference [54] presented a survey that mainly focused on various deep learning techniques based on classification, statistical, and clustering. Additionally, the review also briefly introduced 11 benchmark video datasets and proposed three models to discover inconsistency for anomaly detection. In addition, Reference [60] published a general review of the deep architectures for video anomaly detection using unsupervised and semi-supervised deep learning methods. The goal of the survey divided the state-of-the-art methods into three models including representation learning for reconstruction, predictive modelling, and deep generative models. Similarly, a detailed review of abnormal behaviour recognition in surveillance videos is provided by Reference [90]. This presented different techniques for behaviour representation including features extraction and semantic information about the human action to determine whether the behaviour is normal or not. In addition, the survey presented frameworks and classified abnormal behaviour detection methods both in crowded and uncrowded scenes covering six popular datasets. Moreover, Reference [17] classified types of anomalies and provided different deep anomaly detection techniques. Additionally, the review also discussed the adoption of these methods for real-world problems. Likewise, a survey of violence detection techniques from surveillance videos was introduced by Reference [126]. The aim of this survey provided three types of violence detection methods based on traditional machine learning, **support vector machine (SVM)**, and deep learning. Moreover, video features for violence detection and eight available datasets are also presented. Then, the study by Reference [125] mainly focused on single-scene video anomaly detection. The survey classified thematic grouping by representation (e.g., hand-crafted features, deep learning features) and modelling strategies such as object detection and tracking approach, supervised anomaly detection, and video-level weak supervision. Furthermore, the review also provided various approaches for a single-scene video to detect abnormal activities such as distance-based, probabilistic, and reconstruction-based methods, and summarized five single-scene video datasets. Likewise, References [105, 149] conducted a review of deep learning-based methods for anomaly detection in videos. The purpose of the review provided various frameworks for training and learning based on supervised, unsupervised, semi-supervised, and active learning methods and discussed the performance of these methods covering available benchmark datasets. A survey of Reference [136] presented various anomaly detection approaches in road traffic, focusing mainly on entities such as vehicles, pedestrians, environment. The aim of the review emphasized visual scene learning methods related to anomaly detection as supervised, unsupervised or semi-supervised. Furthermore, anomaly detection approaches (e.g., model-based, proximity-based, classification-based, prediction-based, reconstruction-based) were also presented. Most recently, Reference [113] reviewed deep anomaly detection. This work aimed to provide a categorization of deep anomaly detection and presented complexities and unsolved anomaly detection challenges such as rarity, class imbalance, and diverse types of anomalies. In addition, various deep learning methods for feature representations including generic feature learning and anomaly measure-dependent feature learning were provided.

Concerning anomaly analysis in images, there were several surveys for anomaly problems dealing with images. A survey on image forgery detection was conducted by Reference [144]. The aim of this survey primarily focused on various forgery detection techniques applied on copy-move in digital images and analyzed the advantages and the limitations of each technique. In other surveys by References [12, 94, 120], the classification of various image forgery detection methods were reviewed, i.e., emphasizing on passive or blind techniques, for example, copy-move, and splicing retouching. References [92, 96] reviewed four main types of forgery detection techniques, namely, image splicing, copy-move, re-sampling, and retouching. Similarly, a survey by Reference [31] classified different methods based on five groups including probabilistic novelty detection, distance-based methods, reconstruction-based methods, domain-based methods, and information-theoretic

methods emerging for the background model. Likewise, another survey by Reference [132] focused on different methods for image forgery detection including digital watermarking, digital signature, copy-move, image retouching, and splicing. Most recently, a survey was compiled by Reference [35] concentrated on the types of data used in medicine such as X-ray radiography, **Computed Tomography (CT)** scan. Moreover, the review also presented algorithmic approaches for medical anomaly detection various methods including unsupervised anomaly detection (e.g., autoencoders, generative adversarial networks) and supervised anomaly detection (e.g., multi-task learning, long short-term memory, recurrent neural networks).

1.4 Survey Organization

The remainder of the survey is organized as follows. In Section 2, we introduce four problem descriptions, namely, anomaly detection, anomaly classification, anomaly prediction, and anomaly localization. Section 3 then proposes various techniques for feature extraction and modeling in both images and videos. In Section 4, the benchmark datasets in both images and videos, the evaluation, and the outcomes obtained from different methods of anomaly problems are presented. In Section 5, we discuss the application domains for anomaly problems along with the limitations and challenges for the research outlook. Finally, Section 6 is the conclusion of the article.

2 PROBLEM DESCRIPTIONS

Abnormality has been defined in a number of ways (e.g., unusual behaviour that is different from the norm in real life, statistical infrequency). It depends on the situation, the context, and the lack of uniform norms due to the complexity and vagueness in defining an anomaly/outlier in a variety of real-world domains such as video surveillance [119], financial transactions [181], defect segmentation [117], medical imaging [89], manufacturing industry [9, 177], quality control [101], and cyber-security [139]. There are several definitions of outliers recommended in the previous works. To be more specific, *anomaly* means the occurrence of events or behaviors that are unusual, irregular, unexpected, and unpredictable and thus different from existing patterns [18, 95, 133, 134]. In addition, abnormality [23, 141] means any suspicious activity or any activity in which possibility of happening is very low. It is well known that, in practice, activities such as chaotic activities, traffic rule violations, fighting, riots, burglary, and stampede are considered anomalous ones because of the rare occurrence of these activities in the real world. For example, a person who runs in the street is considered a normal activity. However, this running activity is considered abnormal if she/he runs in a crowded airport. Thus, the mentioned concepts do not entirely capture all of the possible definitions involved anomaly activities or events in real-life situations, but these notions are what researchers have been considering for the past years and motivating new solutions to the problems.

Furthermore, to prevent abnormal activities (e.g., criminal behaviours, crowd violence) in both outdoor and indoor places such as offices, airports, shopping malls, departmental stores, public places, and railway stations, various techniques, and approaches related to action recognition and anomaly analysis in images and videos are proposed to enhance the human safety of public lives and assets [3, 60, 78, 79, 90, 93, 125, 126, 135, 148, 159, 163]. As an example, it is difficult to detect anomalous events in the crowded scenes of real-world surveillance videos, because the detection of abnormal events can be discriminated against as global abnormal events and local abnormal events. In more detail, when the behaviour of the group in the global scene is abnormal, it is referred to as a global abnormal event whereas the behaviour of an individual member is different from their neighbour's behaviour, the local anomalous event is addressed. Motivated by a number of the above-mentioned studies and based on the major types of computer vision tasks in images and videos (e.g., object recognition, object detection, image classification, object positioning, and

Table 3. Summary of Some Important Works and Datasets for Each Problem Category

Dataset	Problem category	Work
UCSD [72], UMN [122], ShanghaiTech [86], Street Scene [123]	Anomaly detection	[37, 40, 79, 86]
UCF-Crime [150], ChestX−ray8 [168], NEU [147]	Anomaly classification	[61, 104, 167, 168]
NYC [2], Pittsburgh [47], SIMCD Prediction [7]	Anomaly prediction	[7, 45, 170]
UCSD [72], UMN [122], MTD [46], MVTec AD [9]	Anomaly localization	[27, 98, 189, 198]

image segmentation), we conduct to classify the problems of abnormal human activities into four groups: (1) Anomaly detection; (2) Anomaly classification; (3) Anomaly prediction; and (4) Anomaly localization. Furthermore, some vital works and datasets for each problem category are also enumerated in Table 3.

2.1 Anomaly Detection

Anomaly detection is a crucial area of real-time monitoring among other research areas of computer vision, because it focused on the automation of the surveillance system and images. Anomaly detection means identifying abnormal activities. The pressing need for the detection and identification of abnormal actions or events is in high demand to reduce or prevent dangerous actions or events that can cause damage to public security. It can be combined and used in various research fields such as human action recognition, automated surveillance system, tracking, and person re-identification. Therefore, in recent years, detecting anomalous or unusual activities from images and videos is an ongoing challenge and a long-standing problem in the computer vision community due to its pervasive applications. Generally, irregular actions or events rarely occur in a confined space than normal activities such as illegal activities, traffic accidents, crimes, fighting, riots, burglary, and shoplifting. These activities should be detected for the safety of people. Moreover, many hurdles in anomaly detection are noisy environment, illumination changes, deformation, and occlusion.

2.2 Anomaly Classification

Anomaly classification has been extensively studied and achieved promising results in computer vision research. This problem aims to use typical actions or event recognition methods that require the whole observation of activities and then extract features and build a model to classify the normal or abnormal activities. Although there have been many studies on anomaly classification in images and videos over the past few years, it remains quite challenging to implement the results for real applications. It is not easy to gain an exact analysis of activities in the image or video due to the diversified backgrounds, angle of photography, and low image or video resolution. Therefore, there are many reasons why this task is still an open problem.

2.3 Anomaly Prediction

Anomaly prediction is an active research topic in computer vision and has a variety of real-world applications, including video surveillance, video retrieval, and the prevention of dangerous events. The significant difference between anomaly detection and anomaly prediction is that the whole actions or events are observed in recognition, while only the beginning actions or events segment is provided in the prediction problem due to the partially observed video. Hence, the ability to predict an abnormal and suspicious activity before it is fully executed is vital in many real-world applications to prevent criminal behaviours, violent incidents, and traffic accidents. There is a great demand for an intelligent surveillance system to detect abnormal events in images and videos. Additionally, it is one of the most challenging to find out the rare events in the crowded or diversity

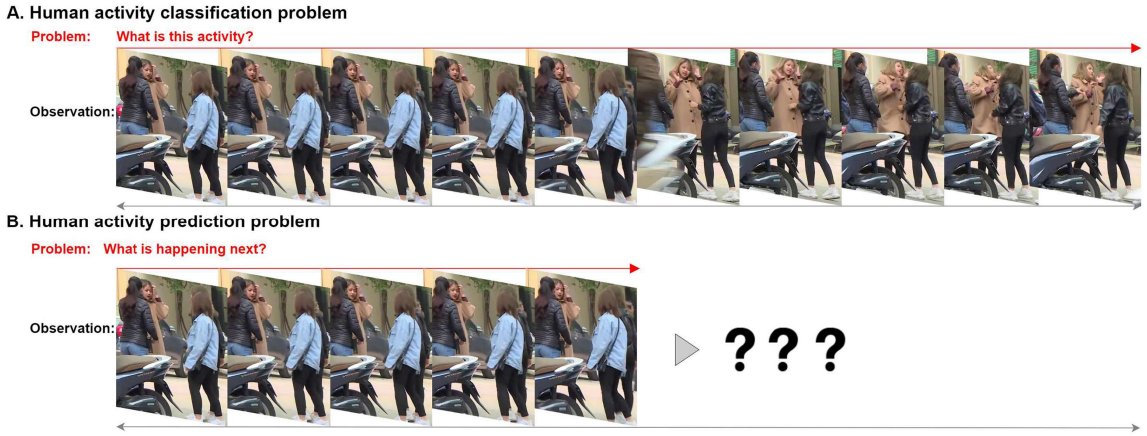


Fig. 1. The difference in anomaly analysis tasks, namely, activity classification and activity prediction [3, 70, 93, 131, 165].

in scenes of videos due to the lack of information when only a fraction of the unusual events are observed. In addition, the difference between activity classification and activity prediction is illustrated in Figure 1. The problem of activity prediction is defined as a probabilistic process of inferring ongoing activities from videos only containing the beginning part of the activities (i.e., unfinished activities provided temporally in incomplete videos). More specifically, future video frames are anticipated based on previous video frames for early prediction of actions or events before they are observed.

2.4 Anomaly Localization

Anomaly localization is a technique that identifies the anomalous region of input images or frames at the pixel level. It is a more complex task that assigns each pixel, or each patch of pixels, an anomaly score to output an anomaly map. Moreover, This problem is another challenging task in videos due to the diversity of possible actions or events and changes drastically over the environment. Here, only a local region in the video is highlighted with an anomaly, for example, an ambulance crossing road irrespective of a traffic signal. Hence, it is imperative to analyze the behaviour of the moving objects to determine whether the action or activity is normal or abnormal. Note that local anomalous behaviour corresponds to the behaviour of a group of objects in a localized region that is different from that of their neighbours in various times and places. Therefore, this topic has piqued the interest of researchers and is particularly significant in the industrial field, where it can be used to automatically identify defective products in images as well as localize human behavioural irregularity in videos. Moreover, several benchmark datasets, such as UCSD [72] and UMN [122], are applicable to both anomaly detection and anomaly localization problems.

3 COMPUTATIONAL MODELS

In this section, we describe different techniques for action recognition problems in images and videos. Furthermore, we also dwell on approaches to feature extraction based on handcrafted features and various approaches for anomaly analysis tasks in the field of deep learning related to images and videos for above mentioned related problems.

3.1 Generic Action Analysis

Various approaches using different frameworks have been proposed over the years for generic action recognition from still images. Different techniques focused on exploiting human pose and

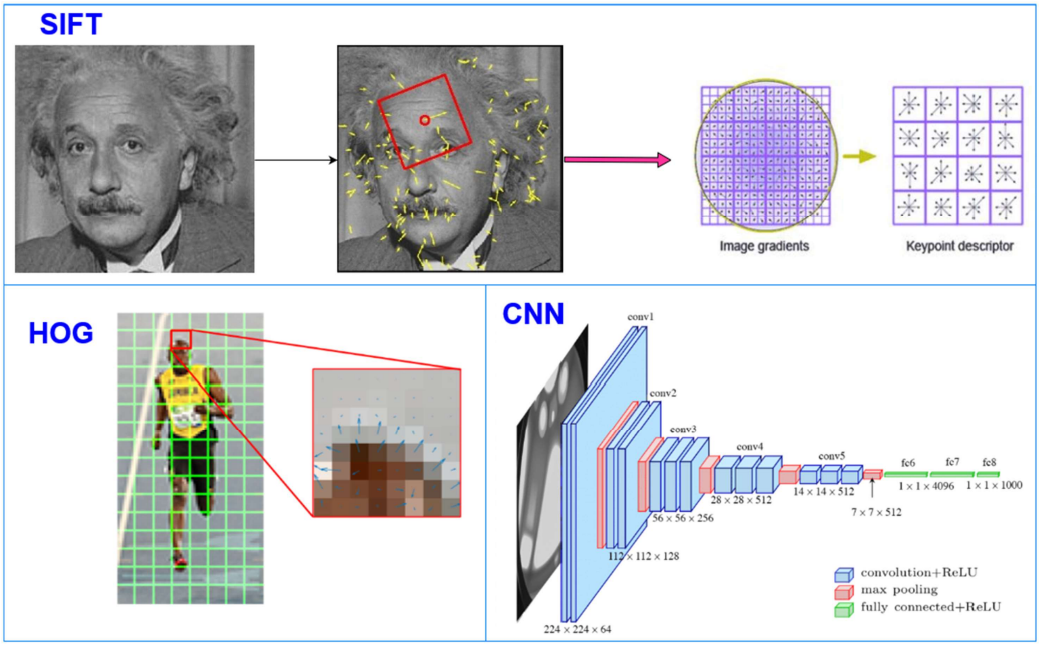


Fig. 2. Notable methods for action recognition in images [25, 83, 138].

context information to recognize actions were proposed by various works [174, 192]. Likewise, an approach of Reference [58] proposed the **Auto-Encoding Variational Bayes (AEVB)** algorithm for image action recognition based on the Stochastic Gradient Variational Bayes estimator via inference and learning simple ancestral sampling to optimize the recognition model. Similarly, Reference [55] proposed an approach based on incorporating colour features into a part-based detection framework for image action detection. In recent years, a lot of approaches for action recognition in images were proposed by various works such as References [4, 19, 116]. These works primarily introduced an architecture based on a **generative adversarial network (GAN)** for image anomaly detection. Furthermore, we also provide some remarkable methods for image action recognition; see Figure 2.

However, regarding the action recognition in videos, Reference [51] introduced an approach for multi-view action recognition based on self-similarities of action sequences that captured the structure of temporal similarities and dissimilarities within an action frame. Next, the approach of References [109, 164, 178] exploited trajectories to extract motion and features for human action recognition. In another work, Reference [173] introduced a method to recognize human actions based on position differences of body joints, called as EigenJoints, and then performed the Naïve-Bayes nearest neighbour [13] as the classifier for action recognition. Similar to Reference [173], the proposed approach of Reference [15, 76, 158] also recognized human action using body-part movements. Likewise, the single-stream and two-stream networks were applied for video action recognition by References [33, 53, 145, 199] respectively. In Reference [145], the temporal network trained on multi-frame dense optical flow to recognize motion while the spatial network performed action recognition from video frames. Next, the proposed approach of Reference [154] used **3D Convolution Net (C3D)**, which is known as 3D ConvNets to learn spatial appearance directly in end-to-end training. Then, Reference [110] proposed fusing handcrafted features and deep learned features to improve the accuracy. Later, Reference [14] proposed state-of-the-art architecture, namely, Two-Stream **Inflated 3D ConvNet (I3D)** based on 2D ConvNet inflation to learn seamless Spatio-temporal feature extractors in videos. Then, the approach of Reference [156],

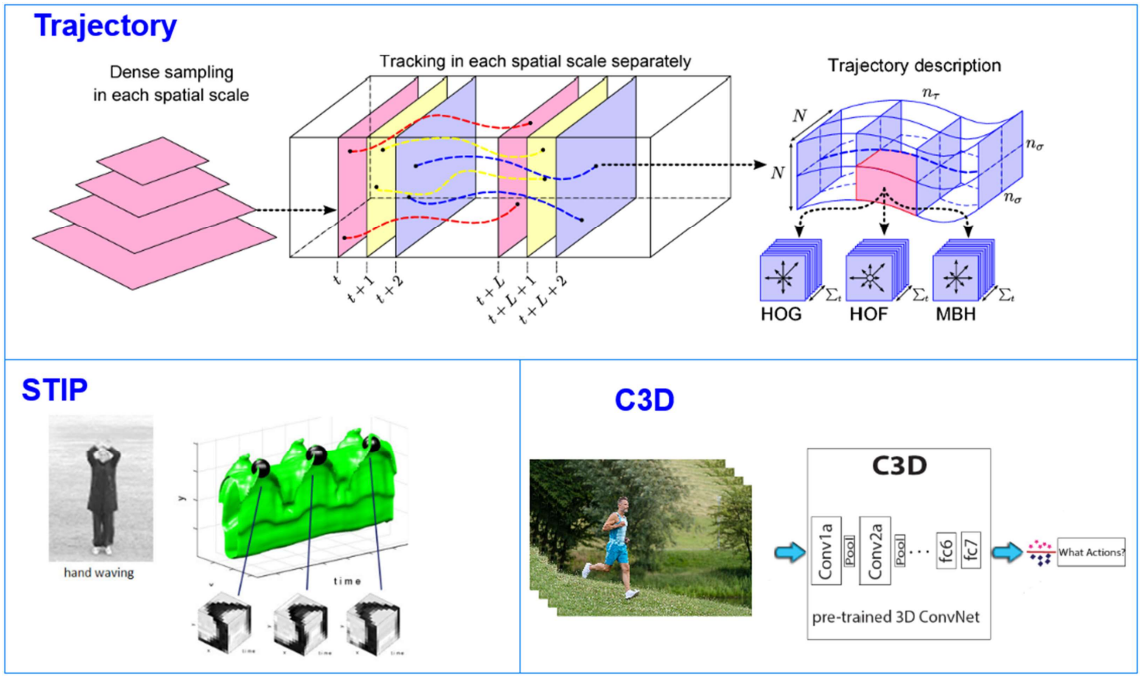


Fig. 3. Notable methods for action recognition in videos [16, 155, 164].

made use of **channel separated convolution networks (CSN)** for video action recognition. The network captured spatial and Spatio-temporal features in their distinct layers to process each direction and then fused them locally at all stages of convolution. Most recently, Reference [193] introduced a framework called **Dynamic Sampling Networks (DSN)** for video action recognition composed of two modules, namely, the sampling module and classification module. In this framework, the sample module selected the most discriminative and relevant clip from each section and then they feed into the classification module to predict the action results of each selected clip. Later, the approach of Reference [20] used a **shift graph convolutional network (Shift-GCN)** to investigate human body skeletons for action recognition. This framework consisted of shift graph operations for spatial and temporal skeleton graphs and lightweight point-wise convolutions to recognize human actions. Similar to References [20, 115] designed a **graph convolutional networks (GCN)** by leveraging neural architecture search. The framework exploited the spatial-temporal correlations between nodes and built a search space in the GCN with multiple dynamic graph modules based on human skeletons for action recognition. Similarly, several techniques were introduced to skeleton-based action recognition with GCN and achieved many encouraging results [121, 142, 190]. Furthermore, an emerging list of prominent approaches for the task of action recognition in videos is presented in Figure 3.

3.2 Anomaly Analysis in Images and Videos

As mentioned before, the analysis of anomalous structures in image data, as well as the analysis of abnormal activities or events in videos is a research endeavour of great interest in the fields of machine learning and computer vision. Furthermore, there has been a surge of interest in developing deep learning approaches for anomaly problems such as anomaly detection, anomaly classification, anomaly prediction, and anomaly localization. In more detail, the increasing problem complexity and rare occurrences of anomalies in a variety of real-world applications lead to data imbalance. This requires specialized solutions to handle data imbalance and unlabelled

data. Thus, anomaly problems in images focus solely on unsupervised learning, self-supervised learning, and semi-supervised learning, whereas similar problems in videos mainly concentrate on unsupervised learning and weakly supervised learning. In this section, we, therefore, aim to present a unifying view that connects traditional shallow and novel deep learning approaches for anomaly analysis tasks in images and videos with the aforementioned approaches. Particularly, we briefly group various techniques and approaches for anomaly analysis tasks in images and videos into two categories based on deep learning revolution: (1) Pre-deep learning methods with handcrafted features; and (2) Deep learning-based methods. The state-of-the-art techniques concentrated on handcrafted features in images and videos are unable to be end-to-end trainable and limited real-time capabilities, because their features need to extract and preprocess and then fed into early deep architectures to classify depending on different tasks. However, deep learning approaches are becoming increasingly popular because of their impressive preprocessing effect on end-to-end trainability and real-time capability. In addition, these approaches have empirically demonstrated significant success in image and video tasks such as object classification, action recognition, image caption, semantic segmentation, anomaly detection, anomaly prediction. More specifically, the usage of convolution layers, pooling layers, batch normalization, fully connected layers, and residual connections in deep networks have been borrowed from the 2D space and applied in the 3D environment with remarkable success in recent years.

3.2.1 Pre-Deep Learning Methods with Handcrafted Features. Traditional video and image processing approaches mainly focus on feature extraction by artificially constructing feature operators and implementing anomaly discrimination.

Hand-crafted features approaches are primarily focused on three modules: (1) extracting features; (2) learning a model to describe the distribution of normal situations or encode normal patterns; (3) identifying the isolated clusters or outliers as anomaly activities. For feature extraction module, various techniques such as **SIFT (Scale-Invariant Feature Transform)** [82], **HOG (Histograms of Oriented Gradients)** [25], **STIP (Space-Time Interest Points)** [67], **HOF (Histograms of Optical Flow)** [26], **MBH (Motion Boundary Histogram)** [26], Dense trajectory [164], Cuboid [131], Motion detection [44], interesting objects tracking and behavior analysis [152] are widely used. HOG focused on static appearance information, whereas HOF captured the local motion information. MBH computed for the horizontal and vertical components of the optical flow and eliminated most texture information from the static background. Some methods are applied for extracting features such as Actionlet [70] and Poselet [127]. However, these methods are not robust in diverse or crowded scenes with multiple occlusions and shadows, and it lacks semantic understanding of scenes and the split of moving targets into pieces. Moreover, to extract spatial-temporal features, Reference [186] adapted a **Markov Random Field (MRF)** for modelling the usual patterns. Reference [1] described the local HOF by an exponential distribution for detection of some types of abnormal events. Next, Reference [56] proposed a space-time MRF for modelling the local optical flow pattern with a mixture of probabilistic principal component analyzers. Then, Reference [91] fitted a Gaussian mixture model to a mixture of dynamic textures and outliers for anomaly detection in crowded scenes. Furthermore, to extract motion features, various approaches based on dense trajectories were proposed by References [88, 111, 118]. These methods captured the trajectory information from each frame and incorporated it with a dense optical flow field to track densely sampled points.

However, the major drawback is that the models are prone to noisy motions such as camera movements. In addition, they heavily rely on tracking, thus the accuracy is significantly reduced in complex scenes. Some of them are time-consuming with high computational complexity, thus it is difficult to satisfy real-time requirements in surveillance videos.

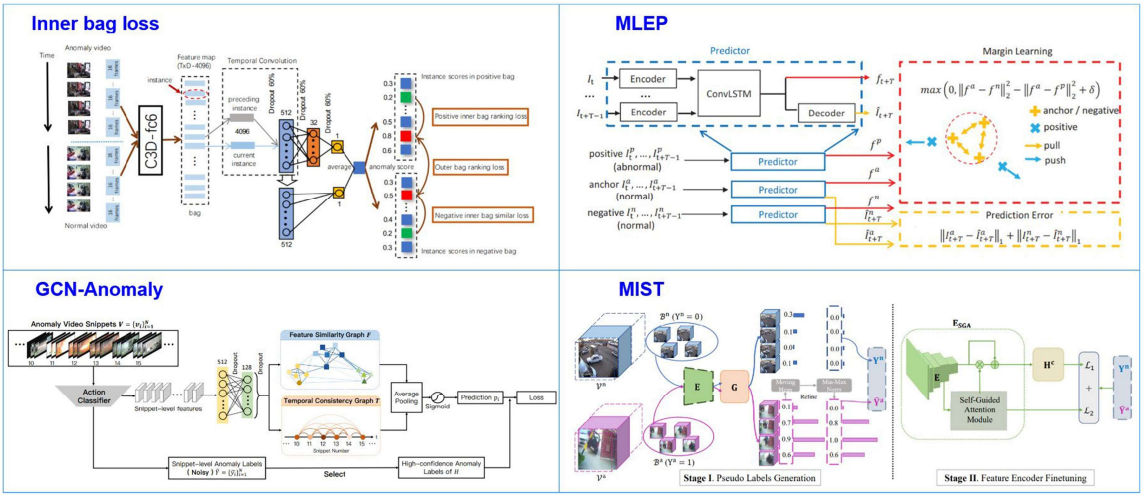


Fig. 4. Notable methods for video anomaly analysis based on weakly supervised deep learning [34, 78, 188, 194].

3.2.2 Deep Learning-based Methods. As mentioned above, anomaly analysis in images and videos (e.g., anomaly detection, anomaly prediction, anomaly localization) is a challenging task due to many reasons: first, anomalies are usually rare so collecting real anomalies for training is often hard or even impossible. Second, the definition of an anomaly does not possess fixed semantics and may refer to different activities or events as well as defective products in images depending on various contexts leading to extracting robust features for modelling anomalies directly unrealistic. Last, anomalies are boundless, or unpredictable in the real world. Thus, the widely used approaches based on deep learning to solve anomaly problems in surveillance videos are unsupervised learning or weakly supervised learning techniques. Furthermore, Figure 4 shows weakly-supervised learning methods while Figure 5 illustrates some state-of-the-art approaches based on unsupervised deep learning for anomaly analysis in videos. In this section, we classify these works into different methods, including (1) Features of Pre-trained Convolutional Neural Networks; (2) Deep Learning Classification Models; (3) Deep Learning Generative Models; (4) Deep Convolutional Autoencoder Models; and (5) Deep Learning Hybrid Models. However, we do not exclude that other taxonomies may also be possible in future works. Furthermore, the taxonomy of methods of anomaly analysis in images and videos are illustrated in Figures 6 and 7, respectively.

3.2.2.1 Features of Pre-trained Convolutional Neural Networks. There exist several methods that use feature descriptors obtained from **Convolutional Neural Networks (CNNs)** that have been pre-trained for anomaly analysis tasks in images and videos. More specifically, Reference [79] proposed a future video frame prediction-based anomaly detection method. The proposed framework identified abnormal events by comparing training data with their expectation instead of reconstructing training data for anomaly detection and predicting the future frame based on its historical observation. Additionally, U-Net [129] network was adopted as a generator to predict the next frame, and a pre-trained network, namely, Flownet [30] was used to estimate optical flow. Moreover, Reference [146] designed an approach based on **Aggregation of Ensembles (AOE)** for anomaly detection in crowd videos where the AOE is an assemble of pre-trained CNNs including AlexNet [63], GoogLeNet [151], and VGGNet [151] to extract higher quality features at different semantic levels from natural images in crowd scenes. In addition, EfficientNet [153] architecture pre-trained on ImageNet [24] was applied as the feature extractor for medical images.

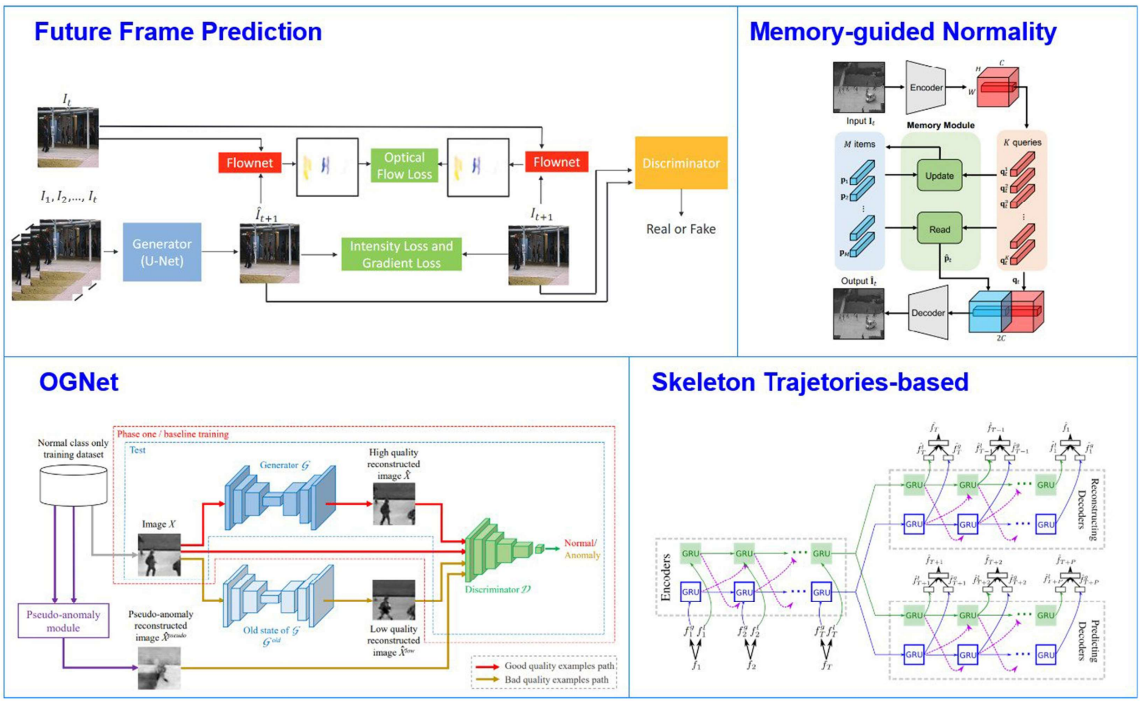


Fig. 5. Notable unsupervised learning methods for anomaly analysis in videos [79, 100, 114, 182].

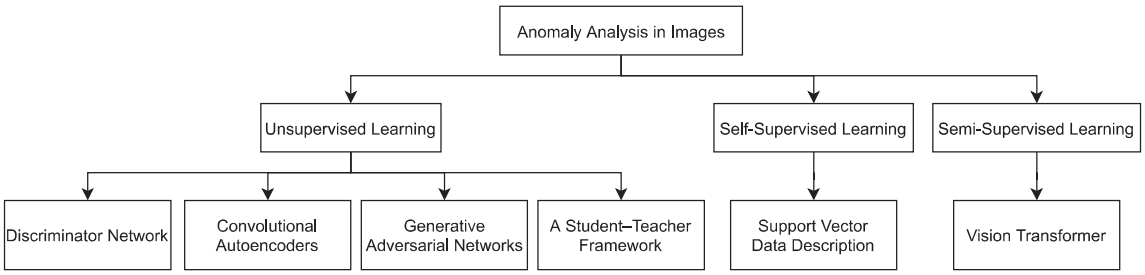


Fig. 6. Taxonomy of methods of anomaly analysis in images.

3.2.2.2 Deep Learning Classification Models. In a work done by Reference [52], an approach for anomalous event detection in crowded scenes by using HOG and the **Histograms of Oriented Swarm Accelerations (HOSA)** to extract appearance and motion features and then applied the One-class SVM [49] framework to detect abnormal events. Next, Reference [172] introduced a method for anomaly detection by using image segmentation with partially occludes target-related RGB data in which Mask-RCNN [41] was implemented to provide semantic segmentation between background and human targets with different experiments. Additionally, depending on each experiment, I3D [14] considering only the RGB stream was applied by using pre-processed data and a one-class SVM [49] with only normal data was applied for the training phase and predicted the abnormality score of testing features. As regards image anomaly detection, Reference [8] proposed a framework to exploit potential anomalies in the training set via addressing the lower-dimensional latent space through a variation of One-class SVM [49] by rejecting the least normal observations. Then, Reference [10] proposed a student-teacher framework for unsupervised anomaly detection in manufacturing images, called Uninformed Students. In this framework, an ensemble of student networks was trained end-to-end on large unlabeled image datasets to mimic the descriptive

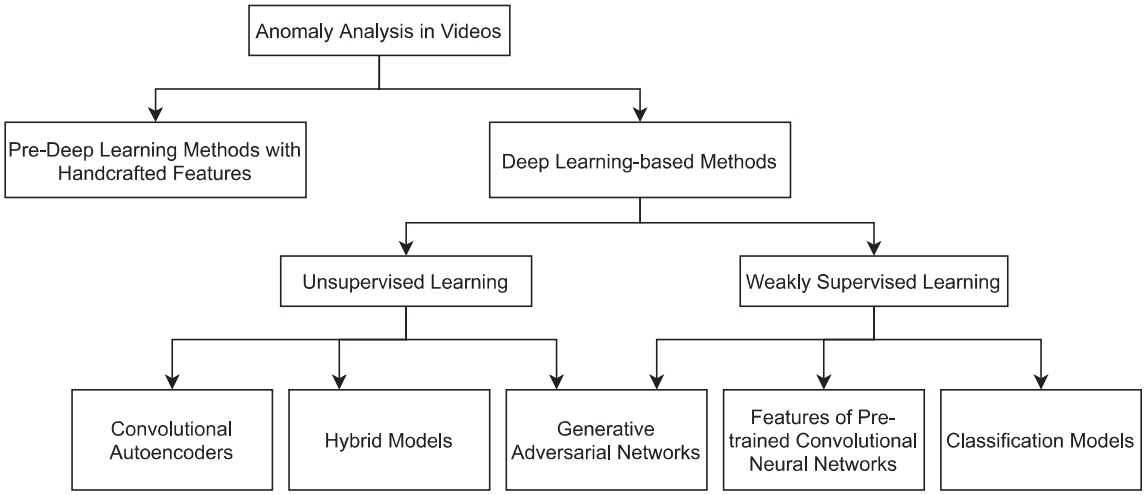


Fig. 7. Taxonomy of methods of anomaly analysis in videos.

teacher network's output wherein local descriptors from a pretrained teacher network serve as surrogate labels for an ensemble of students, and then anomalies were detected when the outputs of the student networks differ from that of the teacher network.

3.2.2.3 Deep Learning Generative Models. Numerous methods for anomaly analysis tasks in images and videos based on GANs, specifically manufacturing images and medical images become more and more popular in these days. In work by Reference [28] designed a dual discriminator-based generative adversarial network GAN [39] structure for video anomaly detection. In the training phase, future frames for normal events were predicted with the generator and then used a frame discriminator and motion discriminator to augment the quality of predicted frames. After that, in the testing phase, the quality of predicted frames and their ground truths were compared to consider those frames with lower prediction qualities as abnormal frames. Next, Reference [184] proposed a GAN-based anomaly detection method, called ALAD via learning an encoder from the data space to the latent space during training wherein normal samples should be accurately reconstructed, whereas anomalous samples will likely be poorly reconstructed. Moreover, the method was also incorporated techniques to improve the encoder network [69] by additional discriminators and stabilized GAN training [99] using spectral normalization. In another one, Reference [137] proposed a fast unsupervised anomaly detection framework based on a GAN for OCT images, namely, f-AnoGAN to identify anomalous images and image segments. The framework consists of two training steps: (1) GAN training on normal images and (2) encoder training based on the trained GAN model. After that, an encoder that maps images to the GAN's latent space for fast inference and anomaly detection via a combined anomaly score based on the building blocks of the trained model. Then, Reference [5] introduced an unsupervised anomaly detection method based on the adversarial training scheme over a skip-connected encoder-decoder network architecture, called Skip-GANomaly. In this framework, skip-connections played a vital role within the generator and feature extraction from the discriminator for the manipulation of hidden features to thoroughly capture the multi-scale distribution of the normal data distribution in high-dimensional image space. Next, Reference [197] introduced a framework called Sparse-GAN for image anomaly detection in retinal disease. The proposed framework includes three modules: (1) image-to-image GAN [50] for medical image anomaly detection; (2) a map of the structured images into latent space [4]; and (3) Sparsity Regularization Net [195]. Recently, a new framework for unsupervised anomaly

detection problem named **Adversarial Predictive coding (APC)** was proposed by Yu et al. [180]. Particularly, this framework included two sequence models, namely, (1) **Recurrent Neural Network (RNN)** and (2) **Gated Recurrent Units (GRU)**, to extract temporal information. Moreover, this framework was also incorporated the prediction framework (GAN) to anticipate the future latent space. Most recently, Reference [32] proposed a method for traffic accident detection in video, namely, SSC-TAD by simultaneously learning the appearance, motion and the context relation consistency within consecutive frames based on a GAN. This framework concentrated on the visual scene context prediction in driving scenarios and the traffic accident detection by considering the temporal frame consistency, temporal object location consistency, and the spatial-temporal relation consistency of road participants from normality to accident situation.

3.2.2.4 Deep Convolutional Autoencoder Models. **Convolutional Autoencoders (CAEs)** [38] are commonly used as a base architecture in unsupervised anomaly detection settings. In more detail, Reference [71] designed a **spatial-temporal cascade autoencoder (ST-CaAE)** based on a classifier for video anomaly detection in crowded scenes. The proposed framework includes two-stream networks, namely, a **spatial-temporal adversarial autoencoder (ST-AAE)** and a **spatial-temporal convolutional autoencoder (ST-CAE)**. Then, Reference [77] proposed a gradient-based visual attention method to explain **variational autoencoders (VAE)** predictions for anomaly localization problems. This method used the learned latent representation to compute gradients and generate visual VAE attention maps and then used them as cues to generate pixel-level binary anomaly masks. Additionally, Reference [85] proposed VAE [59] for image anomaly detection in skin disease with two separate convolution layers for the encoder and avoided using a linear layer to produce mean and log variance. As regards image forgery detection, Reference [57] proposed an approach for robust anomaly detection based on adversarial discriminative parts of images by using the discriminator's **class activation map (CAM)** as a mask for calculating anomaly scores by Grad-CAM [140] from the discriminator network in a Variational Autoencoder [59] with GAN model [39] to visualize the **region of interest (RoI)**. Moreover, SSIM Autoencoder [11] method was applied for unsupervised anomaly detection in manufacturing images.

3.2.2.5 Deep Learning Hybrid Models. There are many research works based on deep learning hybrid models for anomaly analysis in images and videos. To be more specific, Reference [74] proposed a deep neural network model for anomaly detection in videos using spatial-temporal representation learning. In the proposed model, spatial-temporal features were extracted through a multi-scale 3D convolutional neural network and then modelled by a mixed Gaussian model. Additionally, the Mahalanobis distance was calculated to identify anomalous behaviour in different scenes. In addition, Reference [183] introduced an approach based on video-level labels for anomalous event detection. In this method, a batch-based training architecture that learned to maximize scores of the abnormal parts of an input wherein a batch consists of several temporally consecutive segments of a video corresponding to a small portion of a training video instead of a complete video. Next, Reference [78] designed a **Margin Learning Embedded Prediction (MLEP)** framework for open-set supervised video anomaly detection where an open-set setting including some types of anomalies was not contained in the testing set. The proposed framework includes two modules: the video prediction module and the learning margin module. The video prediction module joined a 2D convolution and ConvLSTM [143] to encode motion features and spatial information for future frame prediction while the learning margin module learned to enlarge the gap between normal and abnormal features in the feature space and to decrease the distance between normal features. Then, a temporal encoding network [68] was applied to extract spatial-temporal features of video instances via C3D network [154] and to consider temporal relations between feature instances. Recently, Nasaruddin et al. [103] based on the framework of Sultani et al. [150] via

Table 4. Related Anomalous Image Datasets

Dataset	#Image	Type	Resolution
COIL-100 [107]	7,200	color	640 × 480
CIFAR-100 [62]	60,000	color	32 × 32
ChestX-ray8 [168]	108,948	gray	1024 × 1024
Concrete Crack [112]	40,000	color	277 × 277
MTD [46]	1,344	gray	–
MVTec AD [9]	5,354	gray	700 × 700
		color	1024 × 1024
			600 × 600
BTAD [98]	2,830	color	800 × 600
			1600 × 1600

using the attention mechanism to improve the model robustness for weakly supervised anomaly detection problem. As compared with [150], the attention mechanism was applied to get the foreground of the frames. This addition helped the proposed model achieve significantly higher scores than its baseline.

Concerning anomaly analysis tasks in manufacturing images such as anomaly detection and localization, Reference [27] introduced an anomaly detection and localization approach for Patch Distribution Modeling, named PaDiM based on one-class learning. This method used multivariate Gaussian distributions for patch embedding to get a probabilistic representation of the normal class and then exploited correlations between the different semantic levels of a pre-trained CNN to better localize anomalies. Recently, Reference [189] proposed a method for image anomaly localization based on **successive subspace learning (SSL)** [64, 65], called AnomalyHop. It contains three modules including (1) SSL-based feature extraction to extract features of image patches directly using a data-driven approach; (2) normality feature distributions modelling via Gaussian models to describe the distributions of features of normal images; and (3) anomaly map generation by using the Mahalanobis distance to calculate the anomaly scores and then re-scaled all anomaly maps with the same spatial size and fuse to form the final anomaly map.

4 BENCHMARK DATASETS AND TASK EVALUATIONS

4.1 Related Datasets

4.1.1 Benchmark Datasets for Anomaly Analysis in Images. This section introduces well-known anomaly datasets developed by the researchers from images, as seen in Table 4. Moreover, we briefly summarize anomaly datasets as well as are publicly available for research and useful for the comparison of different methods as follows:

COIL-100 dataset [107] consists of 7,200 color image of 100 objects. The objects were placed on a motorized turntable against the black background and the turntable was rotated 360 degrees to vary the object pose. Images of the objects were taken at pose intervals of 5 degrees with respect to a fixed colour camera and they correspond to 72 poses per object. Each class has 72 images having a resolution of 640 × 480.

CIFAR-100 dataset [62] consists of 100 classes containing 600 images each class. There are 500 training images and 100 test images per class. The 100 classes in this dataset are grouped into 20 superclasses. Each image comes with a "fine" label (the class to which it belongs) and a "coarse" label (the superclass to which it belongs).

ChestX-ray8 dataset [168] includes 108,948 frontal-view X-ray images from 32,717 patients with the text mined eight disease image labels (where each image can have multi-labels), mined from the text radiological reports via natural language processing techniques. In this, a small number of images with pathology are provided with hand-labelled bounding boxes, which can be used as the ground truth to evaluate the disease localization performance. Of 108,948 images, of which 24,636 images contain one or more pathologies, and the remaining 84,312 images are normal cases.

Concrete Crack dataset [112] comprises concrete images, which is divided into two classes, namely, positive and negative crack for image classification. Moreover, each class has 20,000 images with a total of 40,000 color images having a resolution of 277×277.

MTD (Magnetic Tile Defects) dataset [46] comprises 1,344 grayscale images of magnetic tiles under multiple illumination conditions with and without defects. All images with the ROI of magnetic tile are cropped and then classified into six classes including blowhole, crack, fray, break, uneven, and free (no defects). In this dataset, these classes show frayed or uneven areas, cracks, breaks, and blowholes as anomalies. Additionally, a lot of defect-free images contain variations that are similar to anomalies. In addition, a pixel-level label for each defect image is also provided.

MVTec AD dataset [9] comprises 5,354 high-resolution colour images of different objects and textures in the manufacturing industry such as bottles, cables, capsules, metal nuts, and brushes. There are 10 object and 5 texture classes from different domains. Of 5,354 images, 3,629 images are divided into training and 1,725 images for testing. Moreover, it contains grayscale images as well as RGB images and showcases 73 different types of anomalies of different real-world products such as scratches, dents, contamination and different structural changes, on average five per category. All image resolutions are in the range of 700×700 and $1{,}024 \times 1{,}024$ pixels. It is noteworthy that the training set consists of only images without defects whereas the test set contains both images containing various types of defects and defect-free images. In addition, pixel-precise ground truth labels for each defective image region are also provided.

BTAD dataset [98] includes a total of 2,830 real-world images of 3 industrial products showcasing body and surface defects. In this dataset, product 1 has a resolution of $1{,}600 \times 1{,}600$, product 2 is 600×600 and product 3 is 800×600 pixels in size. There are 400 training images for product 1; 1,000 training images for product 2; and 399 train images for product 3, respectively. Furthermore, a pixel-wise ground truth mask for each anomalous image is also given.

4.1.2 Benchmark Datasets for Anomaly Analysis in Videos. This section introduces public benchmark anomaly datasets in videos developed by the researchers from surveillance cameras in different places, as seen in Table 5. To better understand different anomaly datasets, we briefly summarize them as well as are publicly available for research and useful for the comparison of different methods as follows:

UMN dataset [122] contains 11 short videos that are captured in three different scenes: lawn, indoor, and plaza having a resolution of 320×240 and a frame rate of 30 frames per second. The total frames in this dataset are 7,740. It comprises three crowd escaping scenes in both indoor and outdoor scenes. In this dataset, people wandering in groups are normal events, whereas sudden crowds escaping are abnormal events. Furthermore, the frame-level ground truth is provided in the video, which helps to evaluate the performance.

Subway dataset [1] is captured at the entrance and exit gates in a subway station and consists of 1 video. The video is 2 h long in total, with a size 512×384. It represents a realistic scene and contains two categories: Subway Entrance and Subway Exit. The entrance gate video sequence is 1 h 36 min long, whereas the exit gate video footage is 43 min long. In the subway entrance video, the normal activity is people walking, going down the turnstiles, and entering the platform while people walking in the wrong direction, regular interactions between people, running fast and suddenly is also considered as abnormal activities or outliers, and it contains 66 unusual events.

Table 5. Related Anomalous Video Datasets

Dataset	#Action	#Video	Setting	Resolution
UMN [122]	3	11	indoor outdoor	320 × 240
Subway Entrance [1]	66	1	indoor	512 × 384
Subway Exit [1]	19	1	indoor	512 × 384
UCSD Ped1 [72]	40	70	outdoor	238 × 158
UCSD Ped2 [72]	12	28	outdoor	360 × 240
CUHK Avenue [84]	14	37	indoor outdoor	640 × 360
CF-Violence [81]	13	76	outdoor	320 × 240 640 × 480
ShanghaiTech [86]	130	437	outdoor	856 × 480
UCF-Crime [150]	13	1,900	indoor outdoor	320 × 240
Street Scene [123]	205	81	outdoor	1,280 × 720
XD-Violence [169]	6	4,754	indoor outdoor	—

This video contains 144,246 frames (20,000 for training and 124,246 for testing) in total. The Subway Exit surveillance video contains 19 various types of irregular events such as people walking in the wrong direction and loitering near the exit while people exiting from the platform and coming through the turnstiles and turning to the right at the mid of the stairs is a normal activity. The Subway Exit video contains 72,401 frames (7,500 for training and 64,901 for testing) in total.

UCSD dataset [72] is used for anomaly detection problem. It contains video footage of a crowded pedestrian walkway from two various pedestrian scenes: UCSD Ped1 and UCSD Ped2 captured by a static camera. Common anomalies in both these scenes are bikers' movement, small carts, and walking across walkways, while pedestrians were walking along pathways in the normal activity. UCSD Ped1 dataset contains 34 training and 36 testing videos at a low resolution of 238 × 158 pixels whereas the UCSD Ped2 dataset consists of 16 training and 12 testing videos with a higher resolution, 360 × 240 pixels. The UCSD Ped1 set comprises of 14,000 frames (6,800 training, 7,200 testing) in total while UCSD Ped2 set contains 4,560 frames (2,500 training, 2,010 testing) in total. The UCSD dataset provides both frame-level and pixel-level ground truth. The pixel-level ground truth helps to compare localization performance.

CUHK Avenue dataset [84] is captured in CUHK Campus Avenue with 30,652 frames (15,328 training and 15,324 testings) in total with each frame of 640 × 360 resolution and a frame rate of 25 frames per second. Furthermore, the dataset contains 16 training and 21 testing video samples with a total of 14 anomalous events, which include loitering, running, and throwing objects. Additionally, the dataset is challenging for evaluation because of the slight camera shake and the appearance of a few anomalies. In addition, the training videos only have normal events, while testing videos consist of both normal and unusual events. Furthermore, frame level and pixel-level ground truth are provided to evaluate the performance of state-of-the-art anomaly detection methods. However, many outliers are staged so they do not seem natural.

CF-Violence dataset [81] includes violent or non-violent scenes within city center locations from real-life surveillance videos. There are 13 samples of violent behaviour and 63 samples of general behaviour with video resolutions ranging between 320 × 240 and 640 × 480. Additionally,

the violent scenes are classified into two distinct classes of high and low based on the participant population wherein only 4 of the 13 samples can be considered to have a high number of participants.

UCF-Crime dataset [150] contains 1,900 videos including 950 normal videos and 950 unedited videos. Note that the entire video has around 128 h long, with an image dimension of 320 × 240. Furthermore, it contains 13 anomalous activities such as fighting, road accident, burglary, and robbery.

ShanghaiTech dataset [86] comprises of 13 different scenes in campus with 317,398 frames (274,515 training and 42,883 testing) in total. It contains 130 abnormal events captured in complex light conditions and multiple view angles. The unusual activities include bikers, skateboarders, and people fighting. Moreover, the pixel-level ground truth of abnormal events is also annotated both spatially and temporally to evaluate the performance.

Street Scene dataset [123] is captured in Cambridge, MA, with a total of 203,257 video frames (146,410 for training and 56,847 for testing) with a high resolution of 1,280 × 720 pixels. It comprises 46 training video sequences and 35 testing video sequences with a total of 205 anomalous events consisting of 17 different anomaly types, namely, loitering, care outside lane, and jaywalking. These videos are taken from a static USB camera looking down on a scene of a two-lane street with bike lanes and pedestrian sidewalks. Moreover, the dataset is challenging because of changing shadows and moving backgrounds (e.g., a flag and trees blowing in the wind) and the variety of activities taking place, such as cars driving, turning, and parking. Furthermore, the ground truth annotations are provided for each testing video in the form of bounding boxes around each anomalous event in each frame, which helps evaluate the performance.

XD-Violence dataset [169] has a total of 4,754 untrimmed videos with audio signals and weak labels. The dataset consists of 2,405 violent videos and 2,349 non-violent videos and it is collected from multiple sources, such as movies, sports, surveillance, and CCTV, with a total duration of 217 h. The training set contains 3,954 videos with video-level labels, and the test set contains 800 videos including 500 violent videos and 300 non-violent videos. It comprises covers six common types of violence, namely, abuse, car accident, explosion, fighting, riot, and shooting. Moreover, the frame-level ground truth annotations are provided for each testing video to evaluate the performance.

4.2 Evaluations

In the literature on anomaly problems [72, 75, 84, 91], a common evaluation metric is to calculate the **Receiver Operation Characteristic (ROC)** by visualizing the ratio of correctly detected anomalies against incorrectly detected anomalies for varying thresholds and then the **Area Under Curve (AUC)** is cumulated to a scalar for performance evaluation. A higher value indicates better anomaly detection performance. In addition, **Equal Error Rate (EER)** [91] is also calculated as an evaluation metric to evaluate the performance for anomaly analysis. EER means that the percentage of misclassified frames when the **false-positive rate (FPR)** is equal to the false negative rate (miss rate), i.e., FPR = 1 − **true-positive rate (TPR)**. Higher AUC and lower EER mean better performance.

4.2.1 The Evaluation of Methods on Anomalous Image Datasets. In this subsection, we summarize the performance comparison of various benchmark methods on anomalous datasets in images.

Table 6 shows performance comparison of different techniques based on AUC metric on two abnormal image datasets, namely, MVTec AD [9] and MTD [46]. In general, deep learning techniques were high results. As indicated clearly from the MVTec AD dataset, PaDiM method [27] had the highest result at 97.90%, whereas VAE [58] method had the lowest figure at 63.90%. The next most

Table 6. Comparison of Different Methods in Terms of Average AUC on Two Anomalous Image Datasets (%)

Method	AUC	
	MVTec AD [9]	MTD [46]
VAE [58]	63.90	—
GeoTrans [36]	67.20	75.50
GANomaly [4]	76.20	76.60
DSEBM [185]	—	57.20
OC-SVM [6]	71.90	58.70
1-NN [106]	83.90	80.00
Uninformed Students [10]	85.70	—
Patch SVDD [177]	92.10	—
FCDD [80]	92.00	—
VT-ADL [98]	80.70	—
DifferNet [130]	94.90	97.70
PaDiM [27]	97.90	—
AnomalyHop [189]	95.90	—

substantial percentage of AnomalyHop method [189] was at 95.90% and this figure was considerably higher than that of GeoTrans method [36] at 67.20%. Likewise, Patch SVDD method [177] and FCDD method [80] were the same figures at about 92.00%. Additionally, the percentage of the other methods was remarkable results ranging from just nearly 76% to just over 85%. As regards the MTD dataset, it is noticeable that the largest proportion of DifferNet method [130] was at 97.70%, whereas DSEBM [185] had the lowest figure at 57.20%. In addition, this figure of the GeoTrans method [36] was slightly more (about 1%) than that of the GANomaly methods [4] at 75.50% and 76.60%, respectively. Note that 1-NN [106] was far higher than that of the other ones ranging from about 3% to over 20%. From Table 6, we can notice that the results of DifferNet method [130] outperformed on the MTD dataset at nearly 98% but obtained on the MVTec AD dataset at approximately 95%. It is worth noting that the issue of anomaly detection in images is still challenging depending on a specific context.

Our aim in this section is to conduct extensive experiments on two anomalous image datasets, namely, BTAD [98] and MTD [46]. Detailed analysis of the performance and its comparison in terms of ROC-AUC metric at pixel level with three state-of-the-art methodologies including FCDD [80], Patch SVDD [177], and Uninformed Students [10] on these datasets are also reported. To provide a fair comparison, all of the methods here are re-trained and tested with 80% for the training set and 20% for the testing set on MTD dataset similar to experiments have been performed by Rudolph et al. [130]. Note that the testing set contains all abnormal images and some normal images. Furthermore, we use the AUC of the ROC curve measured according to pixel level scores outputted from our experiment to show the performance of three selected methods on two abnormal image datasets. In more detail, Tables 8 and 9 present the results of three methods per class of the BTAD and MTD datasets. Moreover, Figure 8 represents the results of our experiment on the BTAD and MTD datasets in terms of the ROC-AUC metric at the pixel level.

Table 7 shows percentages of three prominent methods, namely, FCDD [80], Patch SVDD [177], and Uninformed Students [10], on two abnormal image datasets, namely, BTAD [98] and MTD [46], in terms of AUC metric at pixel level. Overall, there are substantial differences in results between the state of the art methods on these benchmark datasets. It is obvious that the number of Patch SVDD method is the highest figure of the three prominent methods on the BTAD dataset is at just

Table 7. Comparison of Experimental Methods in Terms of Average AUC Metric at Pixel Level on the BTAD [98] and MTD [46] Datasets (%)

Method	AUC	
	BTAD [98]	MTD [46]
FCDD [80]	84.71	59.94
Patch SVDD [177]	90.04	61.97
Uninformed Students [10]	71.12	67.79

Table 8. Comparison of Experimental Methods in Terms of AUC Metric at Pixel Level on Each Class of BTAD Dataset [98] (%)

Method	AUC		
	Product 1	Product 2	Product 3
FCDD [80]	76.48	85.33	92.32
Patch SVDD [177]	98.44	82.17	89.51
Uninformed Students [10]	68.97	91.00	53.40

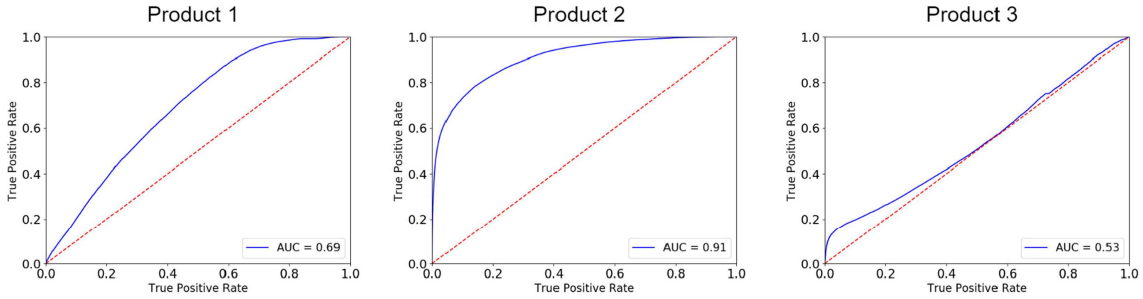
over 90.00% but this method has a significantly low figure on the other dataset at nearly 62.00%. Similar to the Patch SVDD method, the number of FCDD method is far higher than that of the Uninformed Students method on the BTAD dataset at nearly 85.00% and at about 71.00%, respectively. By contrast, the figure of the FCDD method is the lowest at 59.94 %, whereas the number for the Uninformed Students method is the highest at 67.79% about 6% higher, compared to the Patch SVDD method on the MTD dataset. Based on the analysis of these experimental results, it is clear that no the best-performing model in images could achieve performance enhancement of the accuracy and robustness of any anomaly problems specifically.

As can be seen from Table 8, the figures of three prominent methods fluctuate per class of the BTAD dataset, namely, Product 1, Product 2, and Product 3. To be more detailed, the largest percentage of Patch SVDD method [177] on Product 1 of the BTAD dataset is at nearly 98.50% but this method has the lowest figure on Product 2 at just over 82.00%. Moreover, the results of the Uninformed Students method have the lowest for Product 1 and Product 3 at about 69.00% and at roundly 53.50%, respectively. It is notable that, the figure for the FCDD method has twice as many as that of the Uninformed Students method on Product 3. Similarly, Table 9 shows the results of FCDD [80], Patch SVDD [177], and Uninformed Students [10] methods for five class of MTD dataset including Break, Uneven, Fray, Crack, and Blowhole. It is clear that the achieved results of the Uninformed Students method have far higher than the other methods on Crack and Blowhole classes ranging from about 9.00% to nearly 20.00% and have the same figure as the Patch SVDD method on Uneven class at just over 57.00%. However, the percentage of the FCDD method is the highest in the Fray class at approximately 80.00%, compared with 72.89% of the SVDD method and nearly 67.00% of the Uninformed Students method. Likewise, the largest number of Patch SVDD method in Break class is at 72.62%, compared with the other methods in this class, whereas this method has the lowest result in Blowhole class at just over 52.50%. Therefore, the analysis of the detailed figures demonstrates the fact that the results of these prominent methods fluctuate on each type of anomalous dataset in images depending on the realistic conditions and environment.

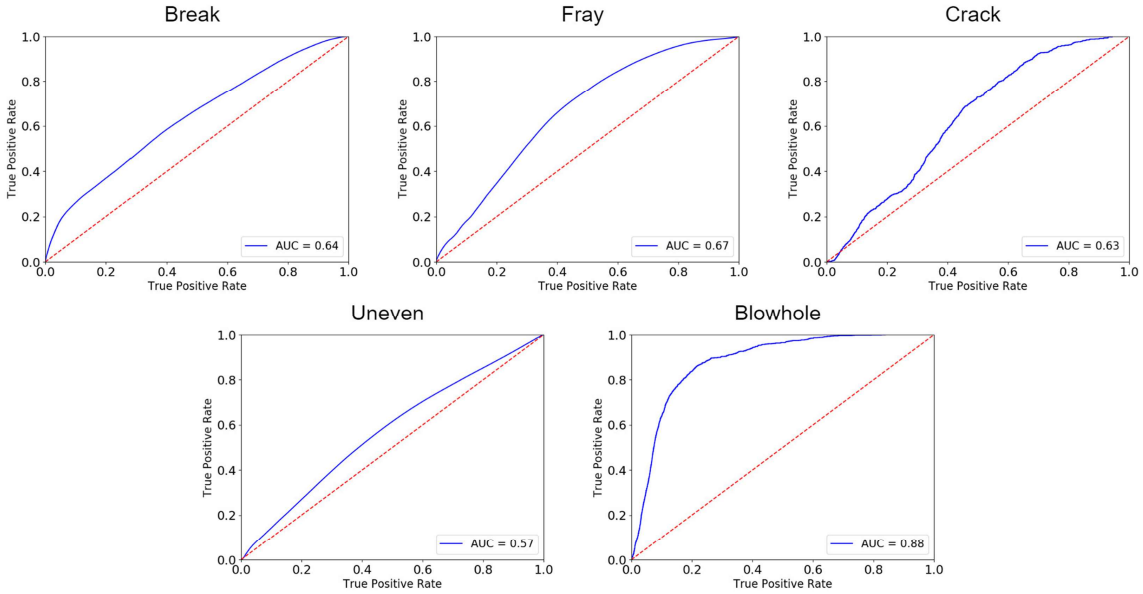
4.2.2 The Evaluation of Methods on Anomalous Video Datasets. In this subsection, we first show some example frames from publicly available datasets, namely, UCSD Ped2, CUHK Avenue

Table 9. Comparison of Experimental Methods in Terms of AUC Metric at Pixel Level per Class of MTD Dataset [46] (%)

Method	AUC				
	Break	Uneven	Fray	Crack	Blowhole
FCDD [80]	58.12	48.22	79.72	54.86	58.78
Patch SVDD [177]	72.62	57.37	72.89	54.44	52.55
Uninformed Students [10]	63.76	57.09	66.81	63.29	88.02



The visualization plot of ROC-AUC score on the BTAD dataset [98].



The visualization plot of ROC-AUC score on the MTD dataset [46].

Fig. 8. Experimental results of Uninformed Students [10] method based on ROC-AUC metric at pixel level per class of the BTAD [98] and the MTD [46] datasets.

(Figure 9). In each dataset, abnormal frames are denoted in red boxes. Additionally, we also summarize the performance comparison of various benchmark methods on these datasets based on AUC and EER metrics, as seen in Table 10.

Table 10 uses AUC and EER metrics at frame level to show a performance comparison of various approaches on two selected anomalous video datasets. Overall, the state-of-the-art techniques based on deep learning performed better than the traditional ones. As regards the UCSD Ped2

Fig. 9. Example images from the UCSD Ped2 [72] and the CUHK Avenue datasets [84]. Top row shows some frames of the UCSD Ped2 dataset, and the second row demonstrates some frames of the CUHK Avenue dataset. Note that red boxes denote anomalies in abnormal frames.

Table 10. Comparison of Different Methods in Terms of AUC and EER Metrics at Frame Level on the UCSD Ped2 [72] and CUHK Avenue [84] Datasets (%)

Method	UCSD Ped2 [72]		CUHK Avenue [84]	
	AUC	EER	AUC	EER
Adam et al. [1]	63.40	42.00	—	—
MPPCA [56]	77.40	30.00	—	—
SF [97]	62.30	42.00	—	—
DTM [91]	84.80	25.00	—	—
MPPCA+SF [91]	71.00	36.00	—	—
Conv-AE [40]	90.00	21.70	70.20	25.10
SL-HOF [167]	95.07	9.00	—	—
Conv-WTA + SVM [157]	92.80	11.20	82.10	24.20
FRCN Action [43]	92.20	13.90	89.80	17.50
RBM [162]	86.43	16.47	78.76	27.21
Spatio-temporal Autoencoder [21]	87.40	12.00	80.30	20.70
Stacked RNN [86]	92.21	—	81.71	—
STAE-grayscale [191]	91.20	16.70	77.10	33.80
STAE-optflow [191]	88.60	20.90	80.90	24.40
Xu et al. [171]	90.80	17.00	—	—
AbnormalGAN [128]	93.50	14.00	—	—
Future Frame Prediction [79]	95.40	—	85.10	—
Li et al. [73]	95.00	6.60	—	—
Narasimhan and Kamath [102]	99.60	16.00	—	—
AnomalyNet [196]	94.90	10.30	86.10	22.00
AnoPCN [175]	96.80	—	86.20	—
Appearance-motion cGAN [108]	96.20	—	86.90	—
MLAD [161]	99.21	2.49	71.54	36.38
MLEP [78]	—	—	92.80	—
Object-centric Auto-encoders [48]	97.80	—	90.40	—
AOE [146]	95.90	—	89.30	—
Doshi and Yilmaz [29]	97.80	—	86.40	—
Lai et al. [66]	95.80	—	87.40	—
Memory-guided Normality [114]	97.00	—	88.50	—
Siamese Distance Learning [124]	94.00	14.10	87.20	18.80
VEC-A [179]	96.90	—	90.20	—
VEC-AM [179]	97.30	—	89.60	—

dataset [72], the percentage of Narasimhan and Kamath [102] based on AUC metric at frame level was the highest at 99.60%, whereas the lowest percentage of SF method [97] was at 62.30%. In addition, the figure for Narasimhan and Kamath [102] was slightly higher than MLAD method [161] with the same metric at 99.60% and 99.21%, respectively. In additionally, the proportion of Doshi and Yilmaz [29] was precisely equal to that of Object-centric auto-encoders method [48] at 97.80%. However, there were substantial differences in the number of EER metrics between Narasimhan

and Kamath [102] and MLAD method at 16.00% and 2.49%, respectively. Moreover, the percentage of SL-HOF method [167] based on the AUC metric was similar to that of Li et al. [73] at about 95.00%. However, the number of SL-HOF methods based on the EER metric was high compared to that of Li et al. [73] at 9.00% and 6.60%, respectively. Regarding the CUHK Avenue dataset [84], the most considerable percentage of MLEP method [78] based on AUC metric at frame level was at 92.80%. Of the other methods, the number of Object-centric Auto-encoders [48] and VEC-A [179] methods were noticeably higher at 90.40% and 90.20%, respectively. By contrast, for the EER metric, the smallest percentage of FRCN Action method [43] accounted for 17.50% and this figure was slightly lower than that of Siamese Distance Learning method [124] at 18.80%.

From the results is presented in Table 10, it is clear that almost all of the state-of-the-art approaches primarily focused on unsupervised or weakly-supervised deep learning are becoming increasingly pervasive on video tasks, specifically anomaly analysis tasks because of the overall performance of these methods outperforming all traditional methods on two different public benchmark datasets by a large margin such as Future Frame Prediction [79], MLEP [78], Memory-guided Normality [114]. It is well known that, in real scenarios, compared with normal events, the anomaly is rare, extremely expensive to collect abnormal events and it is infeasible to collect all possible abnormal events. In addition, the videos contain the complex nature of human behaviours, changing shadows, moving background, diversity of scenes, and view angles. To tackle these problems, two widely used approaches, namely, unsupervised and weakly-supervised learning were applied for anomaly analysis tasks from surveillance videos. In more detail, lots of techniques based on the GAN [39] architecture were applied by adapting U-Net [129] or FlowNet [30] by exploiting regular patterns in terms of appearance and motion on the training set to increase predictive performance. To this end, any pattern that did not agree with these regular ones would be classified as irregular ones. Furthermore, many approaches used an auto-encoder network combined with ConvLSTM [143] and a learning margin module to enlarge the gap between normal and abnormal features in the feature space and to decrease the distance between normal features. In contrast to the above-mentioned approaches, the performance of traditional methods was substantially low ranging from 62.30% to 77.40% such as SF [97], MPPCA [56] due to these methods based on hand-crafted algorithms to learn features and to serve as input for the model to classify abnormal or normal samples. Therefore, these two-stage approaches were computationally expensive, storage demanding, and not end-to-end trainable.

In this section, the goal of our experiments is to evaluate three benchmark methods including Future Frame Prediction [79], MLEP [78], and MNAD [114] on three most commonly used datasets, namely, Street Scene [123], Subway Entrance [1], and UCF-Crime [150] in terms of EER and ROC-AUC metrics at frame-level to evaluate these results of these methods performed on different anomalous video datasets.

We also employ ROC-AUC measured according to frame-level scores outputted from our experiments to indicate the performance of three selected methods on different abnormal video datasets. Figure 10 presents the results of our experiments on the Street Scene, Subway Entrance, and UCF-Crime datasets in terms of the ROC-AUC metric at the frame level. Moreover, the frame-level AUC and EER are listed in Table 11 in which the number of epochs and iterations with the best result of each method showed in Table 12. Furthermore, we provide snapshots of some correct cases and failure cases on two anomalous video datasets from our experiments, as seen in Figures 11 and 12. It is worth noting that the snapshots of incorrect samples indicate some limitations of the anomaly detection models. Regarding the Street Scene [123] dataset, the model is still not robust on the abnormal event relating to the lanes. About the UCF-Crime [150] dataset, the occlusion poses a tremendous challenge for the model to detect anomaly events. In addition, we also provide a

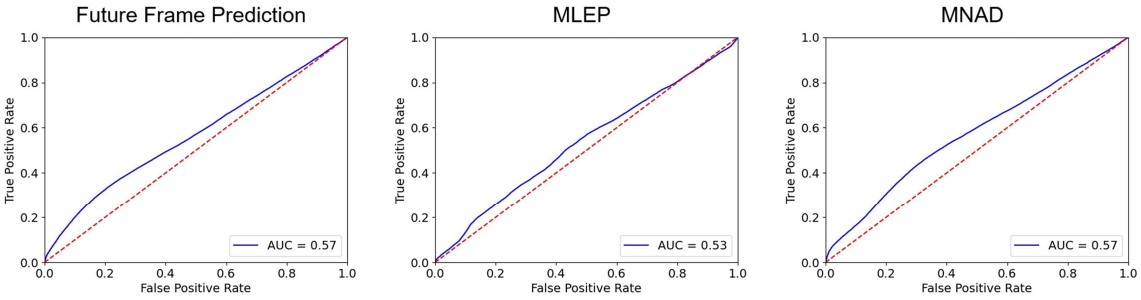
Table 11. Comparison of Experimental Methods in Terms of AUC and EER Metrics at Frame Level on the Street Scene [123], Subway Entrance [1], and UCF-Crime [150] Datasets (%)

Method	Street Scene [123]		Subway Entrance [1]		UCF-Crime [150]	
	AUC	EER	AUC	EER	AUC	EER
Future Frame Prediction [79]	56.53	46.14	71.72	32.22	66.53	38.67
MLEP [78]	53.46	30.49	77.30	46.63	55.08	47.05
MNAD [114]	57.25	44.36	69.37	35.69	65.53	39.69

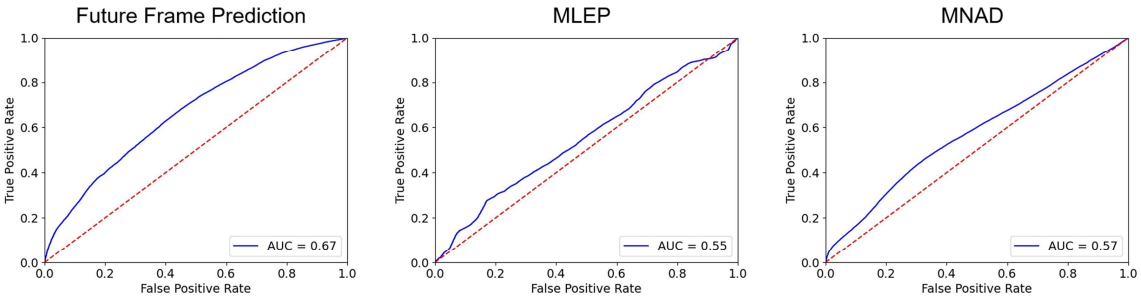
Table 12. Number of Epochs, Iterations That Obtained the Best Result

Dataset	Future Frame Prediction [79]	MLEP [78]	MNAD [114]
Street Scene [123]	3,130	48,000*	37
Subway Entrance [1]	1,000	1,000*	43
UCF-Crime [150]	2,105	10,000*	2

*In MLEP, the training process gets randomly video snippets as input, so we can only get the number of iterations needed to obtain the result.



The visualization plot of ROC-AUC score on the Street Scene dataset [123].



The visualization plot of ROC-AUC score on the UCF-Crime dataset [150].

Fig. 10. Experimental results based on ROC-AUC metric at frame level on the Street Scene [123] and UCF-Crime [150] datasets.

snapshot of Street Scene [123] and UCF-Crime [150] datasets in which abnormal frames are denoted in red boxes and showed in Figure 13.

Table 11 shows percentages of three prominent methods, namely, Future Frame Prediction [79], MLEP [78], and MNAD [114] on three abnormal video datasets in terms of AUC and EER metrics at frame-level. With regard to the AUC metric, the number of MNAD method is the highest figure of the three well-known methods on the Street Scene dataset at 57.25%, however, the figure of this method on the Subway Entrance dataset is the lowest at 69.37%. It is noticeable that there is only

Correct cases for frame prediction on the Street Scene dataset.

Correct cases for frame prediction on the UCF-Crime dataset.

Fig. 11. Correct cases for anomaly prediction on the Street Scene [123], and UCF-Crime [150] datasets.

Failure cases for frame prediction on the Street Scene dataset.

Failure cases for frame prediction on the UCF-Crime dataset.

Fig. 12. Failure cases for anomaly prediction on the Street Scene [123], and UCF-Crime [150] datasets.

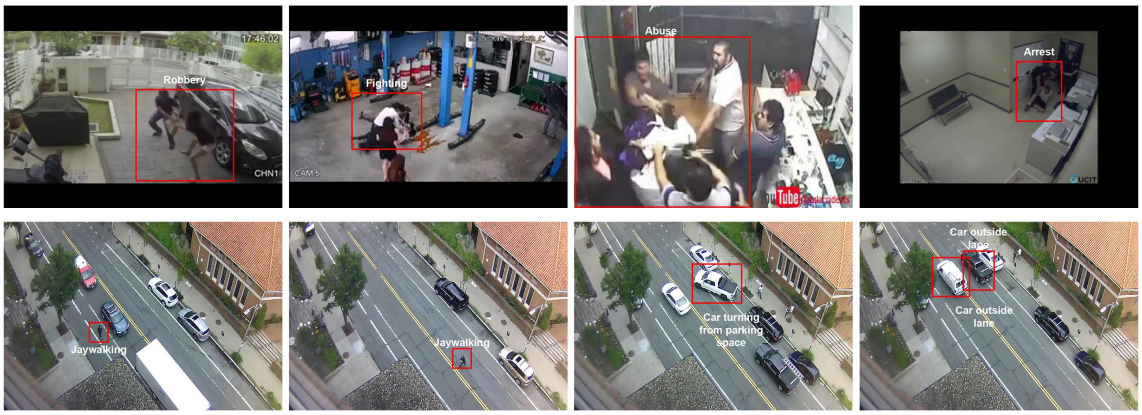


Fig. 13. Image samples from the Street Scene [123] and the UCF-Crime datasets [84]. Rows from top to bottom show: (1) frames of the Street Scene dataset and (2) frames of the UCF-Crime dataset. Red boxes denote anomalies in abnormal frames.

minor variation between the proportion of three methods on the Street Scene dataset, respectively, 56.53% for the Future Frame Prediction method and 53.46% for the MLEP method. Furthermore, the MNAD method is also slightly lower than that of the Future frame prediction method on the UCF-Crime dataset at 65.53% compared to at 66.53%. Nevertheless, the largest percentage of the MLEP method on the Subway Entrance dataset is at 77.30% but this method has the lowest figure on the other datasets at just over 55% and at nearly 53.50%. As regards the EER metric, there are considerable differences in the proportion of these methods on each anomalous video dataset. Of three methods, the percentage of Future Frame Prediction method on the Subway Entrance dataset and the UCF-Crime dataset is the lowest at 32.22% and at 38.67%, respectively, whereas this proportion is the highest on the Street Scene dataset at 46.14%. Similarly, the lowest number of MLEP method on the Street Scene dataset is at nearly 30.50%, however, this method had the

highest figure compared with that of the other ones on the Subway Entrance dataset and the UCF-Crime dataset at 46.63% and 47.05%, respectively. In conclusion, it is clear that there are significant differences in the performance of these benchmark methods depending on each anomalous video dataset. In light of the analysis of these experimental results, it is noteworthy that there are no standard baseline algorithms that could gain efficient accuracy and robustness of any anomaly problems specifically. This means that each baseline framework only achieves high performance of the state-of-the-art depending on a specific context for an anomalous problem in surveillance videos.

5 DISCUSSIONS

As aforementioned, this review of recent advances for anomaly analysis by using image and video processing techniques gives us insightful insights into the current state-of-the-art and possible trends of this application area. In particular, we conduct the analysis of the detailed survey related to anomaly problems from images and videos. Moreover, the taxonomy of anomaly problems for application areas in images and videos is also presented in this article. Furthermore, a significant amount of reported works are investigated based on statistical and filter-based various approaches for anomaly analysis tasks in images and videos. Additionally, a number of real-world datasets and metrics for analyzing anomalies in images and videos are described in detail. In addition, we also summarized and compared a large number of state-of-the-art anomaly problems. Finally, to better understand anomaly analysis from images and videos, we further conducted an empirical assessment of existing state-of-the-art methods on benchmark datasets to demonstrate that the overall performance in images/videos is different depending on anomaly types, movement, illumination, and environmental conditions.

We all know that one of the main challenges in real-world anomaly analysis tasks is the imbalanced data toward normality (i.e., non-anomalous). For this reason, anomalies are typically rare data instances, contrasting to normal instances that often account for an overwhelming proportion of the data. Moreover, abnormal samples are so complex and expensive to collect and there are always unknown and new types of anomalies existing. Therefore, it is difficult to collect a large amount of labelled abnormal instances. This results in the unavailability of large-scale labelled data in most applications. This leads to the fatal limitations of supervised learning for several reasons. First, annotations are time-consuming and require expert annotations who have thorough domain knowledge. Second, even if annotated training corpora are available, the vocabulary of detectable anomalies by a trained model is also limited to those anomaly analysis tasks already known markers for training. Furthermore, there are some challenges for anomaly problems in videos. First, the large number of moving objects in different scenes easily weakens the local anomaly detector. Second, abnormal actions or events only occur for a short period of time leading to difficulty modelling related tasks. Additionally, we also conduct a thorough evaluation of current state-of-the-art unsupervised and weakly unsupervised methods based on deep architectures such as generative adversarial networks, convolutional autoencoders, and feature descriptors using pre-trained convolutional neural networks, as well as traditional methods on publicly available datasets. Even though state-of-the-art methods perform well on the available benchmark datasets, they need a large amount of external training data due to illumination changes, different camera views, diversity in scenes, intra-class and inter-class variation of objects, occlusion of independently moving objects, indoor and outdoor environment, and the lack of comprehensive real-world datasets available for such scenarios.

To tackle the above challenges, the majority of the solutions for anomaly analysis tasks in images/videos rely on unsupervised learning as well as weakly supervised learning. These approaches play an increasingly essential role here, since it is often unknown beforehand what a

variety of real-world anomalies might appear. For the unsupervised learning approach, only normal data are available in the training step, whereas models are trained with both normal data and only a very small amount of anomalous data in the weakly-supervised approach. Note that the unlabeled data, when used in conjunction with a small amount of labelled data, can produce a significant improvement in learning accuracy. It is clear that this has encouraged the development of advanced techniques based on deep learning for a broad range of anomaly analysis tasks in various real-world applications to range from video surveillance, inspection, quality control, financial transactions, manufacturing process monitoring, medical image diagnosis, video surveillance analysis, and so on. In addition to unsupervised learning as well as weakly-unsupervised learning, self-supervised learning is another new research direction with significant potential. Crucially, this approach considers learning from internal cues without requiring labelled data and the frameworks are designed to generate labels automatically. Then, the learned knowledge is transferred to different anomaly analysis tasks in images and videos such as anomaly detection, anomaly prediction, anomaly classification, and anomaly localization.

6 CONCLUSIONS AND FUTURE OUTLOOK

In this article, we conduct the categorization and systematic review of state-of-the-art techniques and approaches for action recognition problems as well as anomaly analysis tasks in images and videos. Moreover, we highlight systematic research on real-world anomaly datasets and metrics related to images and videos. Furthermore, we also conduct extensive experiments and compare the performance of state-of-the-art methods in this research problem. As reviewed, anomaly problems (e.g., anomaly detection, anomaly prediction) are widely studied and applied to a wide range of popular application domains such as marketing, medical diagnosis, network intrusion, fault detection in safety-critical systems, video surveillance, network traffic monitoring, surface defect detection, robotics, and many other applications, not restricted to visual surveillance. To exemplify, one crucial task of anomaly detection or abnormality prediction for surveillance cameras in public or private places such as marketplaces, supermarkets, shopping malls, hospitals, homes, banks, streets, education institutions, city administrative offices, and smart cities is detecting anomalous events such as traffic accidents, crimes, or illegal human activities. Additionally, for instance, the web, online services, and social networks are an essential part of modern life, since they have significantly affected the way people learn, interact in social groups, exchange, store, and search for information. Hence, anomaly detection in the collective behaviour of users is becoming a critical task to detect and track anomalous activity in dynamic networks. In addition, anomalies are very rare events on manufacturing lines. Therefore, anomaly detection automation would enable continuous quality control by reducing human attention and facilitating human operator work.

Regarding future outlook, as mentioned in Sections 4.2.1 and 4.2.2, we observe that the state-of-the-art techniques before the deep learning revolution for tasks like anomaly detection, anomaly prediction mainly focused on hand-crafted features to train data leading to low results of these methods. However, with the deep learning revolution, deep learning architectures prevailed with state-of-the-art performance on image and video tasks because of the effectiveness of deep learning techniques on end-to-end trainability and on real-time capability. It is true that deep learning approaches have empirically demonstrated significant success in learning representations and outperforming traditional approaches. In addition, our experiments show that the existing state-of-the-art approaches based on deep learning still struggle with anomaly due to the ambiguity and diversity of anomalies in various environmental conditions, the complex nature of human behaviors, and the lack of proper datasets.

Furthermore, there is a lack of comprehensive approaches for anomaly problems in the real world due to the rare and unbounded nature of anomalies. To exemplify, the detection of video

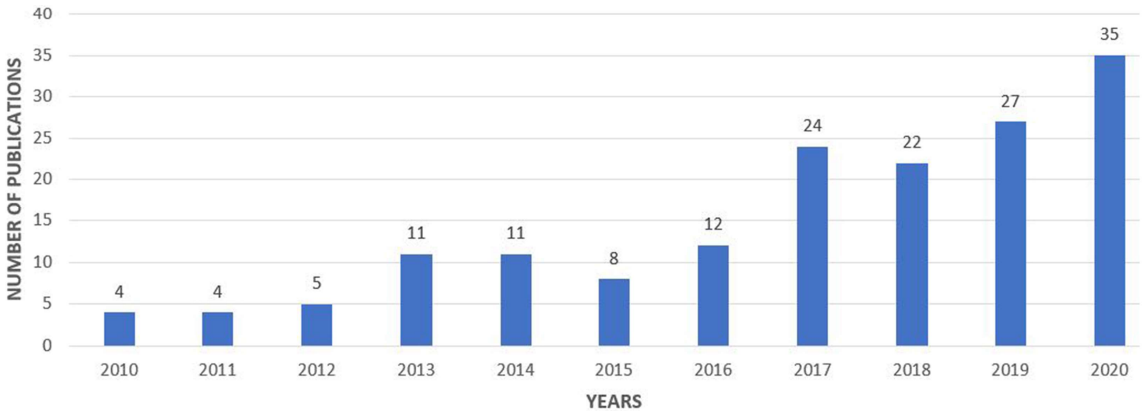


Fig. 14. Number of publications (cited in this comprehensive survey) related to action recognition and anomaly analysis in images and videos over the last decades.

anomalies such as anomalous activities and abnormal entities is challenging due to the ambiguous nature of the anomaly that can be caused by individual actions, a group of activities, complex context with dynamic and time-variant characteristics, various environmental conditions, occlusion of independently moving objects, and variations in appearance. As shown in Figure 14, it should be noted that the number of works for action recognition and anomaly tasks on images or videos increased significantly from 2010 to 2020. Obviously, these tasks are still one of the prominent research directions and need to be addressed in computer vision and machine learning in the future.

In a nutshell, we intensively review research works of anomaly analysis in images and videos. We review a large body of works relating to datasets and methods; and discuss the role of anomaly analysis in the applications. The anomaly analysis research works are categorized into two areas, including images and videos. In addition, we conduct an intensive benchmark of different computational models on popular benchmark datasets. We believe this survey offers a comprehensive overview and suggests important insights for the next generation of research work on anomaly analysis.

REFERENCES

[1] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 3 (2008), 555–560.

[2] NYC Open Data. 2014. Open Data for All New Yorkers. Retrieved from https://opendata.cityofnewyork.us/.

[3] Jake K. Aggarwal and Michael S. Ryoo. 2011. Human activity analysis: A review. *ACM Comput. Surveys* 43, 3 (2011), 1–43.

[4] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Proceedings of the Asian Conference on Computer Vision (ACCV'18)*. 622–637.

[5] Samet Akçay, Amir Atapour-Abarghouei, and Toby P. Breckon. 2019. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'19)*. IEEE, 1–8.

[6] Jerone Andrews, Thomas Tanay, Edward J. Morton, and Lewis D. Griffin. 2016. Transfer representation-learning for anomaly detection. In *Anomaly Detection Workshop in ICML*.

[7] Amna Bamaqa, Mohamed Sedky, Tomasz Bosakowski, Benhur Bakhtiari Bastaki, and Nasser O. Alshammari. 2022. SIMCD: SIMulated crowd data for anomaly detection and prediction. *Expert Syst. Appl.* (2022), 117475.

[8] Laura Beggel, Michael Pfeiffer, and Bernd Bischl. 2019. Robust anomaly detection in images using adversarial autoencoders. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 206–222.

[9] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9592–9600.

[10] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 2020. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4183–4192.

[11] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. 2019. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP'19)*. Scitepress, 372–380.

[12] Charmil Nitin Bharti and Purvi Tandel. 2016. A survey of image forgery detection techniques. In *Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET'16)*. IEEE, 877–881.

[13] Oren Boiman, Eli Shechtman, and Michal Irani. 2008. In defense of nearest-neighbor based image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[14] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[15] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. 2013. Silhouette-based human action recognition using sequences of key poses. *Pattern Recogn. Lett.* 34, 15 (2013), 1799–1807.

[16] Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund, Jordi Gonzalez, and F. Xavier Roca. 2011. A selective spatio-temporal interest point detector for human action recognition in complex scenes. In *Proceedings of the International Conference on Computer Vision*. IEEE, 1776–1783.

[17] Raghavendra Chalapathy and Sanjay Chawla. 2019. Deep learning for anomaly detection: A survey. Retrieved from https://arXiv:1901.03407.

[18] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Comput. Surveys* 41, 3 (2009), 1–58.

[19] Haoqing Cheng, Heng Liu, Fei Gao, and Zhuo Chen. 2020. ADGAN: A scalable GAN-based architecture for image anomaly detection. In *Proceedings of the IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC'20)*, Vol. 1. IEEE, 987–993.

[20] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 183–192.

[21] Yong Shean Chong and Yong Haur Tay. 2017. Abnormal event detection in videos using spatiotemporal autoencoder. In *Proceedings of the International Symposium on Neural Networks*. Springer, 189–196.

[22] Diane Cook, Kyle D. Feuz, and Narayanan C. Krishnan. 2013. Transfer learning for activity recognition: A survey. *Knowl. Info. Syst.* 36, 3 (2013), 537–556.

[23] Peng Cui, Li-Feng Sun, Zhi-Qiang Liu, and Shi-Qiang Yang. 2007. A sequential monte carlo approach to anomaly detection in tracking visual events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[24] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-FCN: Object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. 379–387.

[25] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.

[26] Navneet Dalal, Bill Triggs, and Cordelia Schmid. 2006. Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision*. Springer, 428–441.

[27] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. 2021. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Proceedings of the International Conference on Pattern Recognition*. Springer, 475–489.

[28] Fei Dong, Yu Zhang, and Xiushan Nie. 2020. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access* 8 (2020), 88170–88176.

[29] Keval Doshi and Yasin Yilmaz. 2020. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 254–255.

[30] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2758–2766.

[31] Thibaud Ehret, Axel Davy, Jean-Michel Morel, and Mauricio Delbracio. 2019. Image anomalies: A review and synthesis of detection methods. *J. Math. Imag. Vision* 61, 5 (2019), 710–743.

[32] Jianwu Fang, Jiahuan Qiao, Jie Bai, Hongkai Yu, and Jianru Xue. 2022. Traffic Accident detection via self-supervised consistency learning in driving scenarios. *IEEE Trans. Intell. Transport. Syst.* (2022). https://ieeexplore.ieee.org/document/9733965.

[33] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1933–1941.

[34] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14009–14018.

[35] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2021. Deep learning for medical anomaly detection–A survey. *ACM Comput. Surveys* 54, 7 (2021), 1–37.

[36] Izhak Golan and Ran El-Yaniv. 2018. Deep anomaly detection using geometric transformations. *Advances in Neural Information Processing Systems* 31 (2018), 9758–9769.

[37] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1705–1714.

[38] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT press.

[39] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. Retrieved from https://arXiv:1406.2661.

[40] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K. Roy-Chowdhury, and Larry S. Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 733–742.

[41] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[42] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. 2017. Going deeper into action recognition: A survey. *Image Vision Comput.* 60 (2017), 4–21.

[43] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*. 3619–3627.

[44] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. 2004. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst., Man Cybernet., Part C (Appl. Rev.)* 34, 3 (2004), 334–352.

[45] Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V. Chawla. 2018. DeepCrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1423–1432.

[46] Yibin Huang, Congying Qiu, and Kui Yuan. 2020. Surface defect saliency of magnetic tile. *Visual Comput.* 36, 1 (2020), 85–96.

[47] 311 Response Center in the City of Pittsburgh. 2014. 311 Data. Retrieved from https://data.wprdc.org/dataset/311-data.

[48] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. 2019. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 7842–7851.

[49] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. 2019. Detecting abnormal events in video using narrowed normality clusters. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'19)*. IEEE, 1951–1960.

[50] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1125–1134.

[51] Imran N. Junejo, Emilie Dexter, Ivan Laptev, and Patrick Perez. 2010. View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1 (2010), 172–185.

[52] Vagia Kaltsa, Alexia Briassouli, Ioannis Kompatsiaris, and Michael G. Strintzis. 2014. Swarm-based motion features for anomaly detection in crowds. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'14)*. IEEE, 2353–2357.

[53] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1725–1732.

[54] Phulpreet Kaur, M. Gangadharappa, and Shalu Gautam. 2018. An overview of anomaly detection in video surveillance. In *Proceedings of the International Conference on Advances in Computing, Communication Control and Networking (ICACCCN'18)*. IEEE, 607–614.

[55] Fahad Shahbaz Khan, Rao Muhammad Anwer, Joost Van De Weijer, Andrew D. Bagdanov, Antonio M. Lopez, and Michael Felsberg. 2013. Coloring action recognition in still images. *Int. J. Comput. Vision* 105, 3 (2013), 205–221.

[56] Jaechul Kim and Kristen Grauman. 2009. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2921–2928.

[57] Daiki Kimura, Subhajit Chaudhury, Minori Narita, Asim Munawar, and Ryuki Tachibana. 2020. Adversarial discriminative attention for robust anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2172–2181.
[58] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. Retrieved from https://arXiv:1312.6114.
[59] Diederik P. Kingma and Max Welling. 2014. Stochastic gradient VB and the variational auto-encoder. In *Proceedings of the 2nd International Conference on Learning Representations, ICLR*, Vol. 19.
[60] B. Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. 2018. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *J. Imag.* 4, 2 (2018), 36.
[61] N. Satya Krishna, S. Nagesh Bhattu, D. V. L. N. Somayajulu, N. V. Narendra Kumar, and K. Jaya Shankar Reddy. 2022. GssMILP for anomaly classification in surveillance videos. *Expert Syst. Appl.* (2022), 117451.
[62] Alex Krizhevsky. 2012. Learning multiple layers of features from tiny images. University of Toronto (May 2012).
[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Info. Process. Syst.* 25 (2012), 1097–1105.
[64] C.-C. Jay Kuo. 2016. Understanding convolutional neural networks with a mathematical model. *J. Visual Commun. Image Represent.* 41 (2016), 406–413.
[65] C.-C. Jay Kuo, Min Zhang, Siyang Li, Jiali Duan, and Yueru Chen. 2019. Interpretable convolutional neural networks via feedforward design. *J. Visual Commun. Image Represent.* 60 (2019), 346–359.
[66] Yuandu Lai, Rui Liu, and Yahong Han. 2020. Video anomaly detection via predictive autoencoder with gradient-based attention. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'20)*. IEEE, 1–6.
[67] Ivan Laptev. 2005. On space-time interest points. *Int. J. Comput. Vision* 64, 2 (2005), 107–123.
[68] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 156–165.
[69] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. 2017. Alice: Towards understanding adversarial learning for joint distribution matching. *Adv. Neural Info. Process. Syst.* 30 (2017), 5495–5503.
[70] Kang Li and Yun Fu. 2014. Prediction of human activity by discovering temporal sequence patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 8 (2014), 1644–1657.
[71] Nanjun Li, Faliang Chang, and Chunsheng Liu. 2020. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE Trans. Multimedia* 23 (2020), 203–215.
[72] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 1 (2013), 18–32.
[73] Xiaodan Li, Weihai Li, Bin Liu, Qiankun Liu, and Nenghai Yu. 2018. Object-oriented anomaly detection in surveillance videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. IEEE, 1907–1911.
[74] Zhaoyan Li, Yaoshun Li, and Zhisheng Gao. 2020. Spatiotemporal representation learning for video anomaly detection. *IEEE Access* 8 (2020), 25531–25542.
[75] Benjamin Lindemann, Benjamin Maschler, Nada Sahlab, and Michael Weyrich. 2021. A survey on anomaly detection for technical systems using LSTM networks. *Comput. Industry* 131 (2021), 103498.
[76] Li Liu, Ling Shao, Xiantong Zhen, and Xuelong Li. 2013. Learning discriminative key poses for action recognition. *IEEE Trans. Cybernet.* 43, 6 (2013), 1860–1870.
[77] Wenqian Liu, Runze Li, Meng Zheng, Srikrishna Karanam, Ziyan Wu, Bir Bhanu, Richard J. Radke, and Octavia Camps. 2020. Towards visually explaining variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8642–8651.
[78] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. 2019. Margin learning embedded prediction for video anomaly detection with a few Anomalies. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'19)*. 3023–3030.
[79] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future frame prediction for anomaly detection-a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6536–6545.
[80] Philipp Liznerski, Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. 2020. Explainable deep one-class classification. Retrieved from https://arXiv:2007.01760.
[81] Kaelon Lloyd, Paul L. Rosin, David Marshall, and Simon C. Moore. 2017. Detecting violent and abnormal crowd activity using temporal analysis of grey level co-occurrence matrix (GLCM)-based texture measures. *Mach. Vision Appl.* 28, 3–4 (2017), 361–371.
[82] David G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, Vol. 2. Ieee, 1150–1157.

[83] David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60, 2 (2004), 91–110.
[84] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE International Conference on Computer Vision*. 2720–2727.
[85] Yuchen Lu and Peng Xu. 2018. Anomaly detection for skin disease images using variational autoencoder. Retrieved from https://arXiv:1807.01349.
[86] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*. 341–349.
[87] Junjie Ma, Yaping Dai, and Kaoru Hirota. 2017. A survey of video-based crowd anomaly detection in dense scenes. *J. Adv. Comput. Intell. Intell. Inform.* 21, 2 (2017), 235–246.
[88] Ke Ma, Michael Doescher, and Christopher Bodden. 2015. Anomaly detection in crowded scenes using dense trajectories. University of Wisconsin-Madison.
[89] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recogn.* 110 (2021), 107332.
[90] Amira Ben Mabrouk and Ezzeddine Zagrouba. 2018. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Syst. Appl.* 91 (2018), 480–491.
[91] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1975–1981.
[92] J. Malathi, T. Satya Nagamani, K. N. V. S. K. Vijaya Lakshmi, et al. 2019. Survey: Image forgery and its detection techniques. In *Journal of Physics: Conference Series*, Vol. 1228. IOP Publishing, 012036.
[93] D. Manju and V. Radha. 2018. A survey on human activity prediction techniques. *Int. J. Adv. Technol. Eng. Explor.* 5, 47 (2018), 400–406.
[94] S. Manjunatha and Malini M. Patil. 2017. A survey on image forgery detection techniques. *Dig. Image Process.* 9, 5 (2017), 103–108.
[95] Jefferson Ryan Medel and Andreas Savakis. 2016. Anomaly detection in video using predictive convolutional long short-term memory networks. Retrieved from https://arXiv:1612.00390.
[96] Kunj Bihari Meena and Vipin Tyagi. 2019. Image forgery detection: Survey and future directions. In *Data, Engineering and Applications*. Springer, 163–194.
[97] Ramin Mehran, Alexis Oyama, and Mubarak Shah. 2009. Abnormal crowd behavior detection using social force model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, 935–942.
[98] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. 2021. VT-ADL: A vision transformer network for image anomaly detection and localization. In *IEEE 30th International Symposium on Industrial Electronics (ISIE'21)*. IEEE, 1–6.
[99] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral normalization for generative adversarial networks. Retrieved from https://arXiv:1802.05957.
[100] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. 2019. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11996–12004.
[101] Paolo Napoletano, Flavio Piccoli, and Raimondo Schettini. 2018. Anomaly detection in nanofibrous materials by CNN-based self-similarity. *Sensors* 18, 1 (2018), 209.
[102] Medhini G. Narasimhan and Sowmya Kamath. 2018. Dynamic video anomaly detection and localization using sparse denoising autoencoders. *Multimedia Tools Appl.* 77, 11 (2018), 13173–13195.
[103] Nasaruddin Nasaruddin, Kahlil Muchtar, Afdhal Afdhal, and Alvin Prayuda Juniarta Dwiyantoro. 2020. Deep anomaly detection through visual attention in surveillance videos. *J. Big Data* 7, 1 (2020), 1–17.
[104] Vidhya Natarajan, Tzu-Yi Hung, Sriram Vaikundam, and Liang-Tien Chia. 2017. Convolutional networks for voting-based anomaly classification in metal surface inspection. In *Proceedings of the IEEE International Conference on Industrial Technology (ICIT'17)*. IEEE, 986–991.
[105] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. 2020. A comprehensive review on deep learning-based methods for video anomaly detection. *Image Vision Comput.* (2020), 104078.
[106] Tiago S. Nazare, Rodrigo F. de Mello, and Moacir A. Ponti. 2018. Are pre-trained CNNs good feature extractors for anomaly detection in surveillance videos? Retrieved from https://arXiv:1811.08495.
[107] Sameer A. Nene, Shree K. Nayar, Hiroshi Murase, et al. 1996. Columbia object image library (coil-100). In *Citeseer*.
[108] Trong-Nguyen Nguyen and Jean Meunier. 2019. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1273–1283.
[109] Tam V. Nguyen, Jiashi Feng, and Khang Nguyen. 2018. Denser trajectories of anchor points for action recognition. In *Proceedings of the International Conference on Ubiquitous Information Management and Communication (IMCOM'18)*. 1:1–1:8.

[110] Tam V. Nguyen and Bilal Mirza. 2017. Dual-layer kernel extreme learning machine for action recognition. *Neurocomputing* 260 (2017), 123–130.
[111] Tam V. Nguyen, Zheng Song, and Shuicheng Yan. 2015. STAP: Spatial-temporal attention-aware pooling for action recognition. *IEEE Trans. Circ. Syst. Video Technol.* 25, 1 (2015), 77–86.
[112] Ç. F. Özgenel and A. Gönenç Sorguç. 2018. Performance comparison of pretrained convolutional neural networks on crack detection in buildings. In *Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC'18)*, Vol. 35. IAARC Publications, 1–8.
[113] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM Comput. Surveys* 54, 2 (2021), 1–38.
[114] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14372–14381.
[115] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. 2020. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 34. 2669–2676.
[116] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. 2019. OCGAN: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2898–2906.
[117] Claudio Piciarelli, Danilo Avola, Daniele Pannone, and Gian Luca Foresti. 2018. A vision-based system for internal pipeline inspection. *IEEE Trans. Industr. Info.* 15, 6 (2018), 3289–3299.
[118] Claudio Piciarelli and Gian Luca Foresti. 2006. On-line trajectory clustering for anomalous events detection. *Pattern Recogn. Lett.* 27, 15 (2006), 1835–1842.
[119] Claudio Piciarelli, Christian Micheloni, and Gian Luca Foresti. 2008. Trajectory-based anomalous event detection. *IEEE Trans. Circ. Syst. Video Technol.* 18, 11 (2008), 1544–1554.
[120] Tanzeela Qazi, Khizar Hayat, Samee U. Khan, Sajjad A. Madani, Imran A. Khan, Joanna Kołodziej, Hongxiang Li, Weiyao Lin, Kin Choong Yow, and Cheng-Zhong Xu. 2013. Survey on blind image forgery detection. *IET Image Process.* 7, 7 (2013), 660–670.
[121] Yang Qin, Lingfei Mo, Chenyang Li, and Jiayi Luo. 2020. Skeleton-based action recognition by part-aware graph convolutional networks. *Visual Comput.* 36, 3 (2020), 621–631.
[122] R. Raghavendra, A. D. Bue, and M. Cristani. 2006. Unusual crowd activity dataset of University of Minnesota. http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi.
[123] Bharathkumar Ramachandra and Michael Jones. 2020. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'20)*. 2569–2578.
[124] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. 2020. Learning a distance function with a Siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV'20)*. 2598–2607.
[125] Bharathkumar Ramachandra, Michael Jones, and Ranga Raju Vatsavai. 2022. A Survey of Single-Scene Video Anomaly Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 5 (2022), 2293–2312.
[126] Muhammad Ramzan, Adnan Abid, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Muzamil Ahmed, Mahwish Ilyas, and Ahsan Mahmood. 2019. A review on state-of-the-art violence detection techniques. *IEEE Access* 7 (2019), 107560–107575.
[127] Michalis Raptis and Leonid Sigal. 2013. Poselet key-framing: A model for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2650–2657.
[128] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. 2017. Abnormal event detection in videos using generative adversarial nets. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'17)*. IEEE, 1577–1581.
[129] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 234–241.
[130] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. 2021. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1907–1916.
[131] Michael S. Ryoo. 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proceedings of the International Conference on Computer Vision*. IEEE, 1036–1043.
[132] Akram Hatem Saber, Mohd Ayyub Khan, and Basim Galeb Mejbel. 2020. A survey on image forgery detection using different forensic approaches. *Adv. Sci. Technol. Eng. Syst. J.* 5, 3 (2020), 361–370.

[133] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. 2017. Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Trans. Image Process.* 26, 4 (2017), 1992–2004.
[134] Dinesh Kumar Saini, Dikshika Ahir, and Amit Ganatra. 2016. Techniques and challenges in building intelligent systems: anomaly detection in camera surveillance. In *Proceedings of the 1st International Conference on Information and Communication Technology for Intelligent Systems*. Springer, 11–21.
[135] Sreeja Sankaran Nampoothiri and Anoop Kadan. 2014. Review on vision based human activity analysis. *Int. J. Comput. Appl.* 99 (8 2014), 9–14. https://doi.org/10.5120/17343-6240
[136] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, and Partha Pratim Roy. 2020. Anomaly detection in road traffic using visual surveillance: A survey. *Comput. Surveys* 53, 6 (2020), 1–26.
[137] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. 2019. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Med. Image Anal.* 54 (2019), 30–44.
[138] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Netw.* 61 (2015), 85–117.
[139] Erich Schubert, Arthur Zimek, and Hans-Peter Kriegel. 2014. Local outlier detection reconsidered: A generalized view on locality with applications to spatial, video, and network outlier detection. *Data Min. Knowl. Discov.* 28, 1 (2014), 190–237.
[140] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.
[141] Jean Serra. 2006. A lattice approach to image segmentation. *J. Math. Imag. Vision* 24, 1 (2006), 83–130.
[142] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12026–12035.
[143] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Retrieved from https://arXiv:1506.04214.
[144] B. L. Shivakumar and Lt. Dr. S. Santhosh Baboo. 2010. Detecting copy-move forgery in digital images: A survey and analysis of current methods. *Global Journal of Computer Science and Technology* 10, 7 (2010), 61–65.
[145] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Info. Process. Syst.* 27 (2014).
[146] Kuldeep Singh, Shantanu Rajora, Dinesh Kumar Vishwakarma, Gaurav Tripathi, Sandeep Kumar, and Gurjit Singh Walia. 2020. Crowd anomaly detection using aggregation of ensembles of fine-tuned convnets. *Neurocomputing* 371 (2020), 188–198.
[147] Kechen Song, Shaopeng Hu, and Yunhui Yan. 2014. Automatic recognition of surface defects on hot-rolled steel strip using scattering convolution network. *J. Comput. Info. Syst* 10, 7 (2014), 3049–3055.
[148] Alexandros Stergiou and Ronald Poppe. 2019. Analyzing human–human interactions: A survey. *Comput. Vision Image Understand.* 188 (2019), 102799.
[149] Jessie James P. Suarez and Prospero C. Naval Jr. 2020. A survey on deep learning techniques for video anomaly detection. Retrieved from https://arXiv:2009.14146.
[150] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 6479–6488.
[151] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 1–9.
[152] Ahmed Taha, Hala H. Zayed, M. E. Khalifa, and M. El-sayed. 2014. Exploring behavior analysis in video surveillance applications. *Int. J. Comput. Appl.* 93, 14 (2014).
[153] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*. PMLR, 6105–6114.
[154] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 4489–4497.
[155] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2014. C3D: Generic features for video analysis. Retrieved from https://arxiv.org/abs/1412.0767.
[156] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. 2019. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5552–5561.
[157] Hanh T. M. Tran and David Hogg. 2017. Anomaly detection using a convolutional winner-take-all autoencoder. In *Proceedings of the British Machine Vision Conference*. British Machine Vision Association.

[158] Khai N. Tran, Ioannis A. Kakadiaris, and Shishir K. Shah. 2012. Part-based motion descriptor image for human action recognition. *Pattern Recogn.* 45, 7 (2012), 2562–2572.

[159] Nghia Pham Trong, Hung Nguyen, Kotani Kazunori, and Bac Le Hoai. 2017. A comprehensive survey on human activity prediction. In *Proceedings of the International Conference on Computational Science and Its Applications*. Springer, 411–425.

[160] Vassilios Tsakanikas and Tasos Dagiuklas. 2018. Video surveillance systems-current status and future trends. *Comput. Electric. Eng.* 70 (2018), 736–753.

[161] Hung Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Phung. 2019. Robust anomaly detection in videos using multilevel representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5216–5223.

[162] Hung Vu, Dinh Phung, Tu Dinh Nguyen, Anthony Trevors, and Svetha Venkatesh. 2017. Energy-based models for video anomaly detection. Retrieved from https://arXiv:1708.05211.

[163] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. 2020. Weakly supervised video anomaly detection via center-guided discriminative learning. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'20)*. IEEE, 1–6.

[164] H. Wang, A. Kläser, C. Schmid, and C. Liu. 2011. Action recognition by dense trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. 3169–3176. https://doi.org/10.1109/CVPR.2011.5995407

[165] Haoran Wang, Wankou Yang, Chunfeng Yuan, Haibin Ling, and Weiming Hu. 2017. Human activity prediction using temporally-weighted generalized time warping. *Neurocomputing* 225 (2017), 139–147.

[166] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. 2018. RGB-D-based human motion recognition with deep learning: A survey. *Comput. Vision Image Understand.* 171 (2018), 118–139.

[167] Siqi Wang, En Zhu, Jianping Yin, and Fatih Porikli. 2016. Anomaly detection in crowded scenes by SL-HOF descriptor and foreground classification. In *Proceedings of the 23rd International Conference on Pattern Recognition (ICPR'16)*. IEEE, 3398–3403.

[168] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2097–2106.

[169] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look but also listen: Learning multimodal violence detection under weak supervision. In *Proceedings of the European Conference on Computer Vision*. Springer, 322–339.

[170] Xian Wu, Yuxiao Dong, Chao Huang, Jian Xu, Dong Wang, and Nitesh V. Chawla. 2017. Uapd: Predicting urban anomalies from spatial-temporal data. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 622–638.

[171] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. 2017. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vision Image Understand.* 156 (2017), 117–127.

[172] J. Yan, F. Angelini, and S. M. Naqvi. 2020. Image segmentation based privacy-preserving human action recognition for anomaly detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'20)*. 8931–8935.

[173] Xiaodong Yang and Ying Li Tian. 2012. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 14–19.

[174] Bangpeng Yao and Li Fei-Fei. 2010. Grouplet: A structured image representation for recognizing human and object interactions. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 9–16.

[175] Muchao Ye, Xiaojiang Peng, Weihao Gan, Wei Wu, and Yu Qiao. 2019. Anopcn: Video anomaly detection via deep predictive coding network. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1805–1813.

[176] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. 2013. A survey on human motion analysis from depth data. In *Time-of-flight and Depth Imaging: Sensors, Algorithms, and Applications*. Springer, 149–187.

[177] Jihun Yi and Sungroh Yoon. 2020. Patch svdd: Patch-level SVDD for anomaly detection and segmentation. In *Proceedings of the Asian Conference on Computer Vision*.

[178] Yang Yi and Yikun Lin. 2013. Human action recognition with salient trajectories. *Signal Process.* 93, 11 (2013), 2932–2941.

[179] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. 2020. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*. 583–591.

[180] Jongmin Yu, Jung-Gyun Kim, Jeonghwan Gwak, Byung-Geun Lee, and Moongu Jeon. 2022. Abnormal event detection using adversarial predictive coding for motion and appearance. *Info. Sci.* 586 (2022), 59–73.
[181] Pengfei Yu and Xuesong Yan. 2020. Stock price prediction based on deep neural networks. *Neural Comput. Appl.* 32, 6 (2020), 1609–1628.
[182] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. 2020. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14183–14193.
[183] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. 2020. CLAWS: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. 358–376.
[184] Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. 2018. Adversarially learned anomaly detection. In *Proceedings of the IEEE International conference on data mining (ICDM'18)*. IEEE, 727–736.
[185] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. 2016. Deep structured energy based models for anomaly detection. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1100–1109.
[186] Dong Zhang, Daniel Gatica-Perez, Samy Bengio, and Iain McCowan. 2005. Semi-supervised adapted hmms for unusual event detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 611–618.
[187] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. 2019. A comprehensive survey of vision-based human action recognition methods. *Sensors* 19, 5 (2019), 1005.
[188] Jiangong Zhang, Laiyun Qing, and Jun Miao. 2019. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP'19)*. IEEE, 4030–4034.
[189] Kaitai Zhang, Bin Wang, Wei Wang, Fahad Sohrab, Moncef Gabbouj, and C.-C. Jay Kuo. 2021. Anomalyhop: An ssl-based image anomaly localization method. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 1–5.
[190] Xikun Zhang, Chang Xu, and Dacheng Tao. 2020. Context aware graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 14321–14330. https://doi.org/10.1109/CVPR42600.2020.01434
[191] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. 2017. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*. 1933–1941.
[192] Yin Zheng, Yu-Jin Zhang, Xue Li, and Bao-Di Liu. 2012. Action recognition in still images using a combination of human pose and context information. In *Proceedings of the 19th IEEE International Conference on Image Processing*. IEEE, 785–788.
[193] Yin-Dong Zheng, Zhaoyang Liu, Tong Lu, and Limin Wang. 2020. Dynamic sampling networks for efficient action recognition in videos. *IEEE Trans. Image Process.* 29 (2020), 7970–7983.
[194] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H. Li, and Ge Li. 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1237–1246.
[195] Joey Tianyi Zhou, Kai Di, Jiawei Du, Xi Peng, Hao Yang, Sinno Jialin Pan, Ivor Tsang, Yong Liu, Zheng Qin, and Rick Siow Mong Goh. 2018. Sc2net: Sparse LSTMs for sparse coding. In *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 32.
[196] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. 2019. AnomalyNet: An anomaly detection network for video surveillance. *IEEE Trans. Info. Forensics Secur.* 14, 10 (2019), 2537–2550.
[197] Kang Zhou, Shenghua Gao, Jun Cheng, Zaiwang Gu, Huazhu Fu, Zhi Tu, Jianlong Yang, Yitian Zhao, and Jiang Liu. 2020. Sparse-GAN: Sparsity-constrained generative adversarial network for anomaly detection in retinal OCT image. In *Proceedings of the IEEE 17th International Symposium on Biomedical Imaging (ISBI'20)*. IEEE, 1227–1231.
[198] Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang. 2016. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Process.: Image Commun.* 47 (2016), 358–368.
[199] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. 2018. Hidden two-stream convolutional networks for action recognition. In *Proceedings of the Asian Conference on Computer Vision*. Springer, 363–378.

Received 1 December 2021; revised 5 June 2022; accepted 7 June 2022