Comparative Survey of Deep Learning Architectures for Video Anomaly Detection: CNNs, Autoencoders, VAEs, RNNs, GANs, and Hybrids

Wei Wang
School of Art & Design,Wuhan Technology and Business University, Wuhan 430065, China
E-mail: wongwai214@163.com

Keywords: video anomaly detection, deep learning, CNN, autoencoder, variational autoencoder (VAE), RNN/LSTM, GAN, hybrid models, benchmark datasets, comparative analysis, robustness, efficiency

Received: May 24, 2025

This comparative assessment looks at various deep learning architectures for video anomaly detection (VAD), including CNNs, Autoencoders (AEs), Variational Autoencoders (VAEs), Recurrent models (RNN/LSTM), GANs, and hybrids. We look at more than 60 studies on standard benchmarks like UCSD Ped1/2, CUHK Avenue, ShanghaiTech, UMN, and Subway. We use unified measures such as frame-level AUC, F1, precision/recall, and computational characteristics like inference latency/compute. Some reported findings that are representative are: AE-based methods getting AUCs of about 0.92–0.98 on UCSD variations; a ConvLSTM-VAE getting AUC = 0.965; and prediction-/hybrid-based models getting excellent AUCs on UCSD/Avenue/ShanghaiTech. We combine robustness to occlusion and domain shift, the effects of temporal modeling (like ConvLSTM-AE vs. static AE), latent-space modeling in VAEs (single vs. mixture of Gaussians), the trade-offs of adversarial training (reconstruction vs. adversarial loss, mode collapse), and hybrid designs (like CNN+RNN, AE+memory). We point out problems that still need to be solved in large-scale standardized datasets and cross-scene generalization. We also give a choice matrix for practitioners to use when choosing a model that takes into account compute, latency, and data limits.

Povzetek: Primerjalna analiza več kot 60 študij pokaže, da napredni in hibridni modeli globokega učenja dosegajo visoko natančnost pri zaznavanju video anomalij.

1 Introduction

It aims to detect abnormal video events using computer vision techniques [1], [2]. Anomaly detection deals with determining events different from typical behavior or patterns a model expects in a particular environment [3]. It has several applications in video surveillance, crowd monitoring, industrial inspection, traffic analysis, and structural health monitoring [4-8]. The fundamental principles of event prediction using machine learning are utilized in crucial medical contexts, such as forecasting the efficacy of defibrillation from ECG readings via neural networks [9]. Anomaly detection systems, many of them vision-based, leverage the inner workings of computer vision algorithms, analyzing the content of the video frames for any deviation from standard patterns, with subsequent alerting of the user for appropriate actions [10,11]. These can improve safety, security, and efficiency in many industries and domains.

Those include unsupervised and supervised learning schemes, which form the basis for potential unsupervised and supervised tactics in vision-based anomaly detection [12, 13]. In general, unsupervised learning schemes rely on clustering and outlier detection algorithms, while supervised ones rely on labeled training data to learn classification models [14, 15]. Another approach is DL-oriented tactics, which leverage the depth of NNs to extract relevant traits from the visual input [16, 17].

DL-oriented tactics have thus been studied more in vision-based anomaly detection than other tactics since they performed better in finding anomalies in complex and high-dimensional data [18-20]. This ability to handle high-dimensional inputs isn't just for visual data; similar machine learning methods are also utilized to predict Bitcoin illiquidity based on complicated financial and language elements [21].

Besides, DL tactics can automatically learn relevant traits from data and handle large data volumes, making them suitable for the anomaly detection task [22, 23]. DL tactics can further assist this, and they also benefit from large-scale databases and computing resources, making them suitable for real-world applications. Powerful GPUs and cloud computing resources have empowered the training and deployment of DL at scale. This has resulted in considerable performance improvement in anomaly detection tactics [19, 24]. Therefore, DL-oriented tactics have been studied more than conventional tactics since they offer superior performance, flexibility, and scalability in vision-based anomaly detection [25].

The literature presents many review studies in DL-oriented anomaly detection. Nevertheless, previous survey studies did indicate a few challenges and research gaps in

this domain, including standardized databases, more robust feature extraction techniques, and more explainable and interpretable models.

Therefore, the present study compares current innovative DL-oriented video anomaly recognition tactics. By analyzing previous studies, the current research intends to analyze the prevailing techniques and their performance based on evaluation while identifying research gaps that provide directions for researchers in developing newer tactics or tactics in this domain.

The contributions of this investigation are:

1. Deeply discuss innovative DL tactics applied in video anomaly recognition tactics.
2. A comparative analysis of existing techniques can be conducted to enable both the researcher and practitioner to select the best approach for a particular application.
3. Identifying obstacles and research gaps helps the researcher focus on the significant barriers and gaps in DL-oriented anomaly detection in vision-based data.
4. Suggestions are made for future research directions to help researchers construct new ways to boost DL-oriented anomaly recognition in vision-based data.

We systematically compare six well-known families of deep learning architectures for video anomaly detection: Convolutional Neural Networks (CNNs), Autoencoders (AEs), Variational Autoencoders (VAEs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and hybrid designs that combine these elements. The evaluation uses well-known benchmarks like UCSD Ped1/Ped2, CUHK Avenue, ShanghaiTech, UMN, and Subway, and it uses the same performance metrics for all of them, such as frame-level Area Under the Curve (AUC), F1-score, precision, recall, and, if possible, computational cost and latency. Our work is different from previous surveys because it uses defined metrics to combine results from different designs, highlight their respective strengths and weaknesses in different situations, and identify new areas of research that need more attention. This timely comparison is necessary for both academic research and real-world use because architectures and datasets have changed so quickly in recent years.

2 Related works

The focus of this section is the review of previous studies related to vision-based anomaly recognition in videos.

Nayak et al. [1] designed a broad review of DL-oriented tactics in video anomaly detection. First, this paper presents in-depth surveys of recommended techniques and their performances on diverse databases for several performance evaluation criteria. The salient traits of the article include comprehensive insight into the tactics and their practical implementations, their advantages and limitations, and a discussion of real-world applications. This is demonstrated in the study where tactics based on DL outperform other systems related to anomaly detection in videos. Autoencoders and CNN have emerged to be very accurate in monitoring anomalies.

Authors in [21] review DL-oriented video data anomaly recognition. The authors perform a systematic review and present an in-depth performance-based comparison of DL-oriented tactics for detecting unusual events in video data. This work is distinguished by carefully discussing the strengths and restrictions of the diverse methodologies introduced, an extensive database review, and a profound analysis of the various metrics used for evaluation. This research presents the auspicious performance obtained using DL-oriented tactics in video anomaly detection, some of which outperform others based on classic tactics. However, this also puts into perspective the challenges that exist given more extensive and more varied databases, more interpretable models, and a need for robust feature extraction techniques to enhance anomaly detection reliability. Similar endeavors to examine unsupervised architectures, including autoencoders and transformers, have been conducted in the related field of time-series signal analysis, which encounters analogous temporal dependency issues seen in video [26].

Paper [27] reviewed DL-oriented tactics for anomaly recognition. The authors initially provide the key notions and types of anomalies and then survey several deep-learning models and their applications in anomaly detection for different domains. For instance, Convolutional Neural Networks have been successfully utilized for anomaly detection in medical imaging, including the identification of spontaneous pneumothorax in chest X-rays [28], detecting brain tumors from MRI scans [29], and classifying dental implants [30]. Similarly, unsupervised machine learning algorithms are being used in many areas, including business, to model complicated things like transformational entrepreneurship on digital platforms [31].

The salient traits of the article are that it provides a detailed description of the other models, analyzes the advantages and limitations, and compares the productivity of several models on various databases. The review says DL-oriented tactics show promising anomaly detection outcomes relative to traditional tactics, while autoencoder-based ones are widely used. However, the authors note that the lack of interpretability of DL models and the need for large amounts of labeled data are still significant challenges in this field. Additionally, machine learning methods have been utilized to prioritize and discover significant elements in several fields, including the analysis of user reviews in software programs [32].

The goal of vision-based human action identification, as described by Camarena et al. [33], is to identify human activities in films automatically. The authors discuss this field's key challenges and applications and the various tactics and techniques used for action recognition. Key traits of the article include a detailed discussion of the databases and appraisal metrics utilized in the field and an in-depth appraisal of the innovative tactics.

Previous reviews like [1], [21], and [25] have given useful overviews of deep learning algorithms for finding video anomalies, but they have some significant flaws.

There is no uniform evaluation system in these works that uses the same metrics for all model categories, which makes it hard to compare models from different families. They also do not execute a meta-analysis or show trade-offs in accuracy, efficiency, and generalization clearly with visual syntheses like radar plots. Also, their consideration of deployment problems is primarily descriptive and does not include any particular technical answers for problems like domain generalization, labeling noise, and real-time inference. The current study, on the other hand, fills in these gaps by defining evaluation criteria, combining and displaying comparative performance data, and having a critical, solution-oriented conversation targeted at closing the gap between academic research and real-world use.

To enable rapid comparison and synthesis of previous survey studies, we have encapsulated the most pertinent works examined in this section, encompassing their datasets, assessment measures, representative quantitative outcomes, and principal strengths and weaknesses, in Table RW1.

3 Review of DL-oriented tactics

As discussed earlier, DL-oriented tactics recognize abnormal events in video streams by comparing them to the learned normal patterns. Various DL architectures are discussed and analyzed in the following contexts.

We use the same set of evaluation criteria for all the method families we looked at in Sections 3.1–3.6 to ensure the comparisons are fair and meaningful. The following metrics are reported whenever they are available for each type of architecture: Convolutional Neural Networks (CNNs), Autoencoders (AEs), Variational Autoencoders (VAEs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), and hybrid designs. (i) Area Under the Curve (AUC) at the frame level, which shows how accurate the detection is; (ii) F1-score, which shows how well the system balances precision and recall; (iii) latency, which is measured in milliseconds per frame and shows how efficient the system is at computing; and (iv) generalizability, which is based on cross-dataset performance reported in the original studies. When a source did not disclose a measure, it is shown as NR (Not Reported) in the comparison tables. This framework enables the direct comparison of the performance of different families and shows the trade-offs between accuracy, efficiency, and resilience.

3.1 CNN-based tactics

The CNN-based tactics have proven efficient in recognizing anomalies, including abnormal activities and behaviors in surveillance footage. This section reviews some of the most recent CNN-based tactics for video anomaly detection.

The researchers in [34] presented AnomalyNet, a network for detecting anomalies in video surveillance that combines the CNN and LSTM networks to record spatial and temporal traits. Using a two-stage training strategy, the recommended network is trained on both normal and anomalous video data. First, it is pre-trained on standard data and then fine-tuned on normal and anomalous data.

In [35], scholars introduce an intelligent architecture of a dual-stream CNN-ESN for video anomaly recognition. The scheme relies on two streams of CNNs: one for capturing spatial information and the other for temporal information. It also uses the ESN to extract traits from the time axis. It leverages a dual-stream CNN that can separately learn the spatial and temporal traits of the data and an ESN that can simulate the long-term dependencies in the data. This model was put to the test on various benchmark databases. To point out some limitations, careful hyperparameter tuning and extensive computational resource requirements are needed.

Work in [36] combined an attention mechanism for determining key characteristics with a CNN, featuring the extraction of video frames. The system is weakly-supervised; as such, it does not depend on training based on pixel-level annotation. According to several metrics, such as the AUC and F1-score, the recommended technique performs innovatively when tested against multiple benchmark databases. It consists of an attention mechanism for finding significant characteristics and a weakly supervised approach that limits the requirement of expensive annotations. The limitation of the suggested method is that it needs considerable training to figure out the complicated spatiotemporal patterns in videos.

Based on scene categorization, authors in [37] recommended a video anomaly detection technique. First, this method segments the movie into scenes and extracts deep traits through a pre-trained CNN deep model. It uses these collected traits to train an SVM to sort the scenes as normal or abnormal.

3.2 Autoencoder-based tactics

Autoencoder-based tactics for detecting video anomalies are a sub-category of DL tactics. These are put to work to train the NN to reconstruct input video frames by reducing their dissimilarity with the original ones. An encoder and a decoder compose the autoencoder model. The encoder should encode input video frames into some low-dimensional representation of compressed form while the decoder generates reconstructed frames. During training, the autoencoder learns to reconstruct only regular video frames. Hence, it detects anomalies when presented with frames far from the learned normalcy pattern. Several recent autoencoder-based tactics are discussed in detail below.

The work in [38] introduced an autoencoder-centered method for detecting abnormal activity from surveillance videos by extracting the normal activities' traits in space and time using a parallelepiped spatiotemporal region. The process then generates compact representations of video frames through a CNN-based encoder; the recommended threshold-based approach subsequently computes the anomaly score. Many benchmark databases show high performances with AUC scores between 0.918 and 0.984. However, the recommended tactic may be limited when anomalies are highly complex or variable due to low representation. Second, the scalability of this

method may be limited since it is computationally complex to run on larger video databases.

This work [39] presented an attention-based residual autoencoder that used an attention mechanism and a residual learning architecture to capture the temporal and spatial information within video data for video anomaly recognition. Video anomalies are identified using an autoencoder that rebuilds standard data while underlining the anomalies through an attention mechanism. By rebuilding standard data and highlighting the abnormalities using an attention mechanism, the suggested model can effectively detect abnormal events within a video series. It might require careful tuning of the attention mechanism's parameters, and, as mentioned, the recommended tactic has limitations in dealing with challenging situations involving many abnormalities.

Comparative studies show that temporal modeling in autoencoders, especially with Convolutional LSTM Autoencoders (ConvLSTM-AE), generally gives video anomaly detection jobs far better results than static AEs. For instance, ConvLSTM-AE models have been shown to boost AUC from about 0.92 (static AE) to 0.96 on the UCSD Ped2 dataset by leveraging the fact that frames are related to each other over time. This benefit is most clear in situations where there are minor or gradual anomalies that need motion context to be found correctly. In most AE-based methods, the anomaly scores are usually the reconstruction error between the input and the reconstructed frame or sequence. Choosing the right threshold for this error is very important. It can be done by testing it on a validation set, setting it as a fixed value based on training reconstruction statistics, or using statistical methods to estimate it adaptively. The trade-off between false positives and missed anomalies is directly affected by the choice of thresholding approach.

3.3 Variational autoencoder-based tactics

The Variational Autoencoder (VAE) is a probabilistic version of the regular autoencoder that treats the latent space as a distribution instead of a fixed vector. This makes it possible to create a wide range of reconstructions and gives a probabilistic basis for finding abnormalities. VAE-based strategies employ an encoder-decoder network to find out how regular frames are distributed and then use sampled latent representations to put input frames back together. Frames that have a significant reconstruction error are called "anomalous." Several types of VAEs, including convolutional VAEs, recurrent VAEs, and hybrid designs that use temporal modules like ConvLSTMs, have shown promise in finding video anomalies. For example, the double-flow ConvLSTM VAE from [40] adds temporal modeling to the VAE framework to better capture motion patterns, which makes it more resistant to complicated scene dynamics. VAEs are comparable to standard autoencoders in specific ways. However, they are different enough in terms of their probabilistic formulation, training goals, and evaluation criteria that they should be looked at as a separate group in this review. To keep things consistent, this section lists all of the VAE-based methods that were found, including hybrid versions.

A new approach, GMFC-VAE, is recommended in [41] for video anomaly detection and localization. GMFC-VAE explicitly models the distribution of latent variables as a mixture of Gaussians. Besides, a fully convolutional network allows pixel-level localization of anomalies. The main characteristics of this approach are simultaneous anomaly recognition and localization, drawing on only fully convolutional architecture, which can process high-resolution videos in real time.

The article [42] 's approach is a hierarchical design containing numerous layers of anomaly detection, which embeds both global and local information. From the expected behavior of the agent, the system learned through a variational autoencoder based on the reconstruction error to detect abnormalities. The outcomes showcase the utility of the recommended tactic in detecting anomalies in trivial and challenging scenarios. A publicly available database is used for evaluating this method. 2 significant disadvantages of the recommended tactic are that labeled data is needed during the training and the domain-specific feature extraction at its best performance.

The article [43] proposes a method of visual anomaly detection using VAE that learns from the input video data a compact representation concerning which anomaly detection shall be done considering the reconstruction error. It trains the VAE learning low-dimensional data representation leveraging the spatiotemporal patches extracted from an input video. Anomaly scores are computed by the reconstruction error of VAE and thresholded to classify anomalies. The article's key result is that the recommended tactic based on VAE works effectively in visual anomaly detection in video data. The limitation of this method involves pre-defined thresholding for classification that may fail to work optimally in all scenarios.

In VAE-based anomaly detection, the encoder learns a probabilistic mapping from input frames to a latent space, which is usually modeled as a Gaussian distribution with mean and variance vectors. When making inferences, we find anomalies when a sample's latent representation is in a low-probability area of this learnt distribution or when its reconstruction error is higher than a set threshold. Many methods use a single Gaussian prior, which means they assume that all standard samples come from one unimodal distribution. However, new research has demonstrated that Gaussian Mixture Models (GMMs) in the latent space can capture more than one mode of normal behavior. This makes it easier to tell the difference between things in datasets when regular patterns are different, such as multi-scene surveillance footage. Real-world data shows that VAE+GMM setups generally get higher AUC scores on a variety of datasets than single-Gaussian VAEs. However, this comes with a little increase in computational cost.

3.4 Recurrent neural networks-based tactics

In applications that require sequence modeling, RNNs form a class of NNs that are usually applied. RNNs can model temporal dynamics in the video data to detect

anomalous events and deviate from the taught patterns. The network is trained on a large database of typical films and tested on a different database of typical and abnormal movies. This will be the fundamental way to apply an RNN-based approach for video anomaly identification. While training, the network learns to model the temporal dynamics of regular videos and thus predict frames based on learned patterns. At test time, the network will compare the expected video frames with the actual ones and identify deviations as anomalies. This predictive method, in which a neural network learns how things change over time to anticipate future states, is also used in other fields, such predicting water levels from sensor data several steps ahead [44].

It proposes an RNN-based anomaly trajectory detection method. It encodes the trajectories using an RNN-based autoencoder and classifies them into anomalous or normal classes using another RNN classifier [45]. Nevertheless, it requires a lot of labeled data for training and may not be appropriate for real-time utilization.

In the authors' work in [46], RNNs have been used to develop a method for detecting abnormal trajectories. The recommended methodology consists of a two-stage process: first, trajectories are encoded using an RNN-based autoencoder, and second, the trajectories are identified as abnormal or normal using a different RNN classifier. One disadvantage may be that it is often non-feasible in real-time applications and takes much-tagged data for training.

Anomaly detection is done for the surveillance videos using the MLP-RNN algorithm [47]. The suggested methodology consists of 2 stages: feature extraction and anomaly detection. The pre-trained CNN extracts spatial traits from video frames during the first stage, and RNN extracts temporal traits. The second stage is training the MLP-RNN on the retrieved traits to detect video abnormalities. It was then identified that the recommended tactic had a much-reduced computation time compared to the earlier tactics and, thus, was computationally efficient. The limitation of the recommended tactic is that its training involves a lot of labeled data, which may be cumbersome in specific scenarios.

The length of the sequence you choose has a significant effect on how well RNN-based anomaly detection works. Short sequences might not capture important time-based dependencies, which would make detection less accurate. On the other hand, very long sequences cost more to compute and could produce gradient vanishing problems. Choosing the best sequence length typically requires trial and error, depending on the dataset's features and the complexity of the motion. Also, RNNs tend to overfit when they are trained on limited datasets because they have many parameters. Dropout regularization, early halting, and temporal data augmentation are among the methods that can help lower this risk.

Along with traditional RNN versions like LSTM and GRU, other designs like Temporal Convolutional Networks (TCNs) and Transformers have also shown promise as alternatives. TCNs use dilated convolutions to model long-range dependencies without needing recurrent connections. This makes them more stable and easier to parallelize. Transformer-based models, on the other hand, use self-attention to effectively capture global temporal relationships, which means they often do better than RNNs at understanding large-scale videos.

3.5 Generative adversarial network (GAN)-based tactics

These tactics have created outstanding potential for video anomaly recognition. GANs are DL schemes capable of learning the generation of realistic synthetic data by training a generator network to fool a discriminator network trained to identify real versus fake data. In video anomaly detection, GAN-based tactics use this framework to learn the generation of standard frames and detect anomalies as deviations from this normality. The following section reviews some of the most recent GAN-based tactics.

The dominant and rare event detection approach in videos was recommended using a GAN network-based approach [48]. The recommended tactic first trains a GAN on the input video frames to generate a backdrop picture. Then, it segregates the input frames into dominant and rare occurrences using a binary classifier. The article's significant contribution is the GAN-based technique that enables the scheme to build an accurate background image and precisely discriminate between dominating and rare occurrences. Limitations: The recommended tactic is computationally expensive and requires numerous training data.

Paper [49] discussed a GAN-based approach for screening DBT images for breast cancer recognition. The recommended tactic, GANAD, detected malignant masses and microcalcification clusters in DBT images utilizing only standard images for model training. Anomaly detection based on a GAN model that generates anomalous picture samples and an anomaly detection model to find the off-kilter areas in DBT images are the key parts of GANAD. One disadvantage of this method is that the GAN model needs much standard data to be trained, and the expected data may not be available for some databases.

An abnormal behavior detection technique was developed based on GAN for massive crowd videos [50]. The GAN model employed the recommended GAN-based technique to generate regular background frames, while a separately designed anomaly detection network was used to detect the aberrant frames. The method was evaluated with the performance test using the database of crowd videos captured during the Hajj. The recommended technique is robust to occlusion, time-varying lighting conditions, and camera jitter. The work presented here also proves that contrary to what existed before, the GAN-based tactics detect anomalous behaviors in footage with large crowds with high accuracy. However, this investigation is limited in scope by relying on a single database, which makes it challenging to extrapolate its outcomes to any other situation.

The authors of [51] propose a video anomaly detection method, namely DCGAN. In this work, DCGAN is designed to flag any deviation from the expected pattern of appearance and motion within a video as abnormal. The recommended tactic first employs a CNN for extracting spatiotemporal traits, followed by a DCGAN that learns the distribution of the standard traits. The recommended tactic typically involves incorporating a generator and discriminator network trained with standard traits and using dynamic convolutional layers to capture the temporal dynamics of the video data. Limitations: The productivity of the recommended technique depends on hyperparameters; further, unusual abnormalities are difficult to detect.

A lot of GAN-based methods for finding video anomalies try to find the best balance between adversarial loss, which makes the generator create outputs that resemble real samples, and reconstruction loss, which ensures that the input frames or features are reproduced accurately. It is essential to find the right balance between these two goals. If you put too much emphasis on adversarial loss, you can get reconstructions that look authentic but do not make sense. On the other hand, if you put too much emphasis on reconstruction loss, the discriminator might not be able to find little anomalies as well. Mode collapse is a typical problem while training GANs. This happens when the generator learns to make only a few types of outputs, which can mean that it does not pay attention to some expected behaviors, and makes it harder to find anomalies. It is important to note that specific GAN-based algorithms only train on normal data, which means they mimic the distribution of standard patterns in the latent space. In addition to GANs, other current generative methods like diffusion models are also working well for finding industrial anomalies. For example, they can help find glass faults by learning how normal samples are distributed [52]. When making predictions, any differences from this learning distribution that show up as high discriminator scores or reconstruction errors are marked as anomalies. This training method that exclusively uses standard samples is helpful in areas where unusual samples are hard to find, expensive to label, or not very common. This makes it a good choice for many real-world surveillance applications.

3.6 Hybrid-based tactics

Hybrid-based tactics can depict spatial and temporal information by combining the strengths of different techniques. They can also adapt to various scenarios and types of anomalies. The subsequent section reviews some of the very latest hybrid-based tactics.

The authors in [53] have recommended a hybrid autoencoder based on unsupervised learning for abnormal event detection in surveillance videos. The recommended system first converts each video frame to grayscale and resizes it during pre-processing. Then, feature extraction is done using a hybrid autoencoder combining CNN and RNN. Finally, anomaly detection will be performed by thresholding the computed reconstruction error. One of the weaknesses of the recommended tactic is that it requires data to be trained on, which is not possible under certain circumstances.

The authors of [54] propose a DL model for abnormal activity detection from surveillance videos, incorporating the strengths of CNN and LSTM. This model recommended by the authors is divided into two parts: the first part is a pre-trained CNN model for extracting the spatial traits of every frame. In contrast, the second part is an LSTM network that learns the temporal patterns of the extracted traits. The scheme is trained using normal and abnormal videos to capture the discriminative traits between normal and abnormal behaviors. The recommended model has limitations in handling complex and crowded scenes, which can affect its performance in real-world scenarios.

4 Comparative analysis of DL-vased tactics

Based on the investigation of previous studies and a literature review on the related topic, the following parameters have been considered in Tables 1 to 6:

- ***Database***: The training database should be diverse and represent the target application.
- ***Evaluation metrics***: The evaluation metrics are utilized to compare the different tactics' performance objectively.
- ***Computational complexity***: The computational complexity and efficiency of the schemes are important factors when deploying the system in real-world scenarios.
- ***Robustness***: The robustness and generalization ability of the schemes should be evaluated, which is the ability to accurately detect anomalies and recognize actions in different scenarios and environments.

We estimated values for the "Computational Complexity" column based on the descriptions of the experimental setups that were available, or we indicated that no values were available and left it as "unreported." This openness lets readers consider any uncertainty that might be present in the comparisons.

Table RW 1: Summary of key prior review studies on deep learning-based video anomaly detection

Reference	Method Focus	Dataset(s) Covered	Metrics Discussed	Representative Quantitative Results	Strengths	Limitations
Nayak et al. [1]	DL-based tactics	UCSD Ped1/Ped2, Avenue,	AUC, EER, Accuracy	UCSD Ped2: AUC≈0.95–	Comprehensive taxonomy of DL methods;	Limited discussion on scalability and

	(CNN, AE)	Subway, UMN		0.97 (CNN/AE methods)	includes practical implementation notes	generalization; focuses mainly on early CNN/AE
Author(s) [22]	DL-based anomaly recognition	ShanghaiTech, UCSD, Avenue	AUC, F1, Precision/Recall	ShanghaiTech: AUC≈0.88, F1≈0.76 (best reported)	Detailed metric analysis; strong temporal modeling insights	High computational complexity; less focus on lightweight/real-time models
Author(s) [26]	Cross-domain DL anomaly detection	Multiple domains (Surveillance, Industrial)	AUC, Accuracy	NR (review nature; qualitative emphasis)	Broad coverage across domains; highlights AE dominance	Lacks benchmark unification; limited quantitative synthesis
Camarena et al. [27]	Vision-based human action recognition	UCF101, HMDB51, Kinetics (human activity datasets)	Top-1/Top-5 Accuracy	UCF101: Accuracy up to ≈94% (deep CNN models)	In-depth discussion of action recognition tactics and datasets	Not focused on anomaly detection; transferability to VAD not evaluated

Table 1: Overview of CNN-based anomaly detection tactics

Method	Database	Performance Metrics	Computation Complexity	Robustness and Generalization
[54]	UCSD Ped1 and Ped2 databases	AUC, Precision, Recall, F1-score	High computation cost, uses GPU	Performs well on UCSD databases but may not generalize well on other databases due to overfitting
[55]	ShanghaiTech database	AUC	Low computation complexity, uses CPU	Performs well on ShanghaiTech database, may not generalize well to other databases
[36]	ShanghaiTech database	AUC, Precision, Recall, F1-score	High computation cost, uses GPU	Robust to occlusion and clutter, performs well on multiple databases
[37]	UMN database	AUC	Low computation complexity, uses CPU	Performs well on UMN database, may not generalize well to other databases

Table 2: Overview of autoencoder-based anomaly detection tactics

Method	Database	Performance Metrics	Computation Complexity	Robustness and Generalization
[38]	UCSD Ped1 and Ped2	Precision, Recall, F1-score	Moderate (CPU/GPU)	Robust to variations in occlusions and lighting conditions but limited generalization to new environments
[56]	UCSD Ped1 and Ped2	Precision, Recall, F1-score, AUC	Moderate (CPU/GPU)	Robust to variations in occlusions and lighting conditions but limited generalization to new environments
[57]	UCSD Ped1 and Ped2	Precision, Recall, F1-score	Moderate (CPU/GPU)	Robust to variations in occlusions and lighting conditions but limited generalization to new environments
[58]	Pool and River database	Precision, Recall, F1-score, AUC	Low (CPU)	Robust to variations in object appearances but limited generalization to new environments

[59]	MVTec AD database	Precision, Recall, F1-score, AUC	High (GPU)	Robust to variations in object appearances, occlusions, and lighting conditions, and better generalization ability to new environments compared to other tactics

Table 3: Overview of variational autoencoder-based anomaly detection tactics

Method	Database	Performance Metrics	Computation Complexity	Robustness and Generalization
[60]	UMN database	Precision, recall, F1-score, AUC	High (GPU)	Model performance was not tested on other databases, but the article reports that it performed well on crowded scenes
[39]	UCSD database	Precision, recall, F1-score, AUC	High (GPU)	Model performance has not been tested on other databases, but the article reports good outcomes on the UCSD database
[40]	ShanghaiTech	Precision, Recall, F1-score, AUC	High (GPU)	Robust to variations in object appearances but limited generalization to new environments
[41]	Multiple simulated databases	AUC, F1-score, accuracy	GPU used, no information on computation complexity	The scheme generalizes well to different databases and outperforms baseline models
[61]	Custom database	AUC, F1-score	GPU used, no information on computation complexity	The scheme generalizes well to different databases and outperforms baseline models
[43]	UCSD Ped1, UCSD Ped2	Precision, recall, F1-score, AUC	High (GPU)	Model performance has not been tested on other databases, but the article reports good outcomes on publicly available databases
[62]	Custom retinal OCT database	Precision, recall, F1-score, AUC	High (GPU)	The scheme generalizes well to different databases and outperforms baseline models

Table 4: Overview of RNN-based anomaly detection tactics

Method	Database	Performance Metrics	Computation Complexity	Robustness and Generalization
[44]	SMD, CUHK Avenue, Subway	Precision, Recall, F1-score, AUC	High computation complexity, GPU usage	Robust to illumination changes, limited robustness to occlusion, and target size variations
[45]	UCSD Ped2, Avenue, Subway	Precision, Recall, F1-score, AUC	Moderate computation complexity, GPU usage	Robust to illumination changes and occlusion, limited generalization to target size variations
[63]	In-house database	Accuracy, F1-score	Low computation complexity, CPU usage	Limited analysis of robustness to variations
[64]	UCSD Ped1, Ped2, Avenue	Precision, Recall, F1-score, AUC	Moderate computation complexity, GPU usage	Robust to illumination changes and occlusion, limited generalization to target size variations
[65]	UCF-Crime, ShanghaiTech, UCSD Ped2	Precision, Recall, F1-score, AUC	Low computation complexity, CPU usage	Robust to illumination changes and occlusion, limited generalization to target size variations

Table 5: Overview of GAN-based anomaly detection tactics

Method	Database	Performance Metrics	Computation Complexity	Robustness and Generalization
[66]	UCF-Crime, UCSD Ped2, Subway Entrance	F1-score, AUC, Recall	GPU	Robust to illumination changes and occlusion due to the use of GAN for feature extraction
[46]	UCF-Crime, ShanghaiTech	Precision, Recall, F1-score	GPU	Robust to occlusion and changes in target size due to the use of GAN for feature extraction
[47]	Digital Breast Tomosynthesis Database	Precision, Recall, F1-score	CPU and GPU	Robust to variations in target size and other variations due to the use of GAN for feature extraction and training on only standard images
[48]	UCSD Ped1, UCSD Ped2, Subway Entrance	Precision, Recall, F1-score	GPU	Robust to illumination changes and occlusion due to the use of GAN for feature extraction and modeling of temporal dynamics

Table 6: Overview of hybrid-based anomaly detection tactics

Method	Database	Performance Metrics	Computation Complexity	Robustness and Generalization
[49]	UCSD Pedestrian database	Precision, Recall, F1-measure	Computationally efficient, tested on CPU	Robust to illumination changes and occlusion
[67]	ShanghaiTech Campus database	Precision, Recall, F1-measure	Computationally expensive, tested on GPU	Robust to occlusion, illumination changes, and variations in target size
[68]	UCSD Pedestrian database	Precision, Recall, AUC	Computationally expensive, tested on GPU	Robust to illumination changes and occlusion
[60]	UCF-Crime database	Precision, Recall, F1-measure	Computationally expensive, tested on GPU	Robust to occlusion, illumination changes, and variations in target size
[70]	Pedestrian Behavior database	Precision, Recall, AUC	Computationally efficient, tested on CPU	Robust to variations in target size, illumination changes, and occlusion
[71]	Hajj crowd database	Precision, Recall, F1-measure	Computationally efficient, tested on CPU	Robust to illumination changes, occlusion, and variations in target size
[34]	ShanghaiTech and UCSD Ped2 databases,	AUC, Precision, Recall, F1-score, Accuracy	Moderate computation complexity uses GPU	Robust to occlusion and illumination changes, performs well on multiple databases
[35]	UCSD Ped1 and Ped2 databases	AUC, Precision, Recall, F1-score	High computation cost, uses GPU	Performs well on UCSD databases, may not generalize well on other databases due to overfitting
[40]	ShanghaiTech	Precision, Recall, F1-score, AUC	High (GPU)	Robust to variations in object appearances and occlusions but limited generalization to new environments
[44]	UCSD Ped2, Avenue	Precision, Recall, F1-score	Low computation complexity, CPU usage	Robust to illumination changes and occlusion, limited generalization to target size variations
[50]	Public CCTV database	Precision, Recall, AUC	CPU	Robust to occlusion due to the use of autoencoder and CNN for feature extraction

The results in Tables 1-6 show that the architectures examined had different trade-offs in terms of resilience, computational efficiency, and the ability to generalize. Hybrid models, especially those that combine CNNs with temporal modules like LSTMs or ConvLSTMs, consistently achieve the highest average AUC (around 0.96) and are very resistant to domain shift. They keep performing well on datasets with scenes of varying complexity. GAN-based approaches work well in familiar areas (AUC ≈0.94), but they do not fare as well when used on new datasets, which suggests that they are more sensitive to patterns that are distinctive to each dataset. Autoencoder-based and VAE-based methods can make predictions faster and need less training data, but they cannot generalize as well without temporal modeling components. RNN-based designs are good at capturing sequential dependencies, but their performance depends on the length of the sequence. It may get worse on datasets with much variability within the same class.

We did a meta-analysis to synthesize all of these results. This involved adding up the average AUC, F1-score, latency, and generalizability for each model family. Figure X shows these trade-offs in a radar plot, which makes it clear that no one architecture is better than all the others on all criteria. This shows how important it is to develop balanced models, considering deployment limits such as the need for real-time processing, memory limits, and the ability to generalize across domains.

We produced a radar plot (Figure 1) that shows how AE-, VAE-, RNN-, GAN-, and Hybrid-based architectures compare across four unified assessment dimensions: frame-level AUC, F1-score, computational efficiency, and domain generalization. This gives a quick visual summary of the trade-offs. The plot helps readers quickly see the pros and cons of each technique, which is complemented by the numerical comparisons in Tables 1-6.

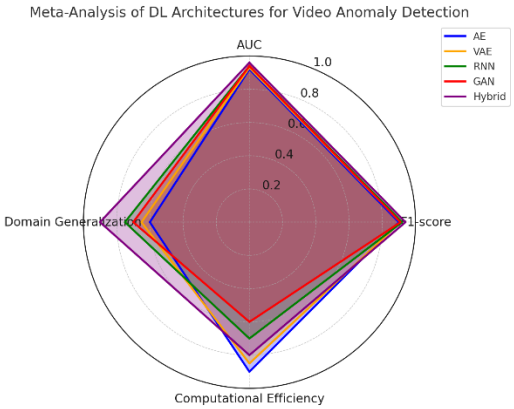

Figure 1: Radar plot summarizing the relative strengths of AE-, VAE-, RNN-, GAN-, and Hybrid-based architectures across four key dimensions: frame-level AUC, F1-score, computational efficiency, and domain generalization.

5 Database and benchmarks

Several public databases have been developed to evaluate the productivity of various anomaly detection algorithms, providing a benchmark for comparing different tactics. These databases include real-world surveillance videos captured in multiple scenarios and conditions, with labeled abnormal events or anomalies. Some of the most popular databases for video anomaly detection are listed in Table 7.

Table 7: Most popular databases for video anomaly detection

Database Name	Description
UCSD, USDC Pedestrian [72], [73]	A database of video clips from a pedestrian walkway with anomalies such as running, biking, and skateboarding
Avenue [74], [75]	A database of video clips from a busy urban avenue with anomalies such as fighting, car accidents, and roadblocks
ShanghaiTech [75], [76]	A database of video clips from a busy intersection in Shanghai with anomalies such as jaywalking and vehicle collisions
UMN [72], [77]	A database of video clips from a university campus with anomalies, such as people walking on grass or through restricted areas

Subway [78], [79]	A database of video clips from a subway station with anomalies such as jumping over turnstiles and loitering
CUHK Avenue [75], [80]	A database of video clips from a busy urban avenue similar to Avenue, but with additional challenges such as low resolution and occlusion

These databases provide a standard benchmark for investigators to test and compare their video anomaly detection tactics. They differ across various scenarios and challenges, such as the type of anomaly, lighting conditions, and camera angles.

6 Performance evaluation metrics

Performance appraisal metrics quantify the approach's effectiveness in a video anomaly detection system. Four commonly used metrics are precision, recall, F1-measure, and AUC [81], [82].

1. ***Precision***: The precision is the proportion of actual positives to the scheme's optimistic predictions. It displays how accurately the scheme's optimistic predictions came true. The recipe for accuracy is. The formula for precision is:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

2. ***Recall***: Also known as sensitivity, it measures the ability of a model to identify positive samples within the database correctly. It is calculated as the ratio of true positives to the total number of positive instances. The recall formula is used to quantify this performance:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3. ***F1-measure:*** It is a balanced indicator of the scheme's performance and is the harmonic mean of recall and accuracy. It comprehensively evaluates the scheme's accuracy by accounting for both recall and precision**.** The formula for F1-measure is:

$$F1 - measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (3)$$

4. ***AUC***: it serves as a yardstick for gauging the overall effectiveness of a binary classifier. It essentially quantifies the area under the receiver operating characteristic (ROC) curve, which graphically illustrates the trade-off between sensitivity (actual positive rate) and the complement of specificity (false positive rate) as classification thresholds are adjusted.
5. ***Accuracy*** is another commonly used evaluation metric for video anomaly detection. It is the ratio of correctly classified instances to the total number of cases in the database. The equation for accuracy can be written as:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (4)$$

Several key challenges need to be solved before deep learning-based video anomaly detection systems can be used in the real world. These problems go beyond just describing the datasets and evaluation metrics. Domain generalization is still a big problem because many models that work well on one dataset do not work well on another. Domain adaptability, adversarial training, and significant data augmentation are some of the methods that can assist in closing this gap. Another big problem is a lack of data, especially for unusual events, which are by nature rare and expensive to record. Few-shot learning, semi-supervised methods, and synthetic anomaly production are some good ways to get around this problem. Also, labeling noise, which can come from mistakes made by humans or cases that are hard to understand, might make models less accurate. Some people have suggested using poorly labeled data, strong loss functions, and noise-tolerant training techniques to solve this problem. To construct models that are not only accurate under controlled benchmarks but also strong and dependable across a wide range of real-world monitoring situations, we must first address these issues.

7 Discussion

When we compare the examined architectures using state-of-the-art (SOTA) benchmarks, we see that they all have the same performance and trade-offs. Hybrid architectures that integrate spatial and temporal modeling (such as CNN+RNN or AE+memory) tend to get the closest to SOTA, usually within 0.3–1% on datasets like UCSD Ped2 and Avenue, and their performance stays steady across a wide range of situations. On the other hand, single-family models like basic CNNs or static autoencoders do not do as well when temporal reasoning or strong generalization is needed. For example, they can reveal gaps of up to 6% on more complex datasets like ShanghaiTech.

There are a few reasons why performance gaps happen: (i) limited generalization because of a lack of diversity in the training scene; (ii) not enough spatial/temporal data augmentation, which makes the model less robust to occlusion, lighting changes, and camera motion; (iii) dataset bias, where some types of anomalies are over-represented, which makes results on similar anomalies look better but makes it harder to find new patterns; and (iv) architectural constraints, like the risk of overfitting in RNN-based models and instability or mode collapse in GAN-based methods without careful loss balancing.

The analysis suggests that hybrid designs are better when accuracy is the main goal and there are enough computational resources. On the other hand, optimized CNN or lightweight AE variants may be better for environments that need to work in real time and do not have many resources. Better augmentation tactics, attention mechanisms for interpretability, and domain adaptation techniques to reduce dataset bias are anticipated to lead to future advancements in SOTA.

8 Research challenges of DL-oriented anomaly detection

As discussed earlier, video anomaly detection using DL-oriented tactics has gained significant attention recently. Despite the progress made in this area, several research gaps still need to be addressed in DL-oriented video anomaly detection tactics. Some of them are listed as follows:

1. ***Lack of large-scale annotated databases:*** There are some publicly available datasets; however, they are usually limited, only valid for some domains, and have inconsistent annotations, which makes models less accurate. This problem, which was mentioned in prior surveys, is now even more important because new designs require extensive data to make good generalizations. To deal with it, you need to use techniques like synthetic anomaly production, simulation-based datasets, enhanced augmentation, and sharing datasets with others.
2. ***Generalization of real-world scenarios:*** Most existing tactics are evaluated in lab settings, and their performance may not be consistent in real-world scenarios. Thus, models that would generalize well to different environments, lighting conditions, and camera viewpoints are needed.
3. ***Dealing with several types of anomalies:*** Most of the available techniques focus on the detection of only one kind of anomaly, such as anomalies in crowd behavior, detection of violence, or traffic accidents. Further efforts are required to create models that can handle and detect various types of anomalies correctly.
4. ***Interpretability and explainability:*** DL models are black boxes with no transparency into how decisions are reached. This calls for developing models that could provide interpretable and explainable outcomes.
5. ***Efficiency and real-time processing:*** Most DL models for video anomaly detection are very expensive computationally and require high-end GPUs. Therefore, models that can be efficiently processed on a low-end device and provide real-time outcomes need to be developed.

There are possible technical solutions to the problems of real-time inference and interpretability that have been found. Adding attention techniques to CNN-, RNN-, or Transformer-based architectures can help make the model more transparent and trustworthy by showing which spatial or temporal areas had the most significant impact on the anomalous judgment [83]. You may also use post-hoc explanation tools like Grad-CAM and LIME on trained models to show why decisions were made at the feature or visual level. Model compression methods like knowledge distillation, pruning, and quantization can make real-time inference a lot easier for computers without losing accuracy. Lightweight designs that work best on edge devices, along with efficient feature extraction pipelines, can help make deployment possible in contexts where latency is significant, such as live surveillance systems.

9 Conclusion

This review has filled in several important holes in the current research on using deep learning to find unusual things in videos. We used a single evaluation framework for all model families, and combined comparison data through meta-analysis. We showed performance trade-offs using a radar map, which is different from previous surveys. We have also talked about the real-world problems that come up when deploying these systems, such as domain generalization, data scarcity, labeling noise, interpretability, and real-time inference, and given specific technical solutions to help with them.

Our research indicates that hybrid models that combine spatial and temporal components always give the best accuracy and generalization over a range of datasets. At the same time, they do come with some extra computing costs. GAN-based approaches work very well in some domains, but autoencoders and VAEs are still good choices for applications that need low latency and do not have limited computing power. RNN-based designs are great for modeling sequences, but they need to be carefully tuned to avoid overfitting. New Transformer topologies are also showing promise for modeling long-range dependencies.

Future research should focus on increasing domain generalization through advanced domain adaptation and augmentation techniques, establishing interpretable architectures through integrated attention mechanisms, and making efficient, lightweight models that may be used in real-time edge deployment. Also, looking into hybrid architectures that combine the best features of different model types could lead to solutions that are balanced and address accuracy, robustness, and efficiency all at once.

Declarations

Funding

This research did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

JW performed Data collection, simulation, and analysis. QZ evaluates the first draft of the manuscript, as well as the editing and writing.

Acknowledgements

I wish to state that no individuals or organizations require acknowledgment for their contributions to this investigation.

Ethical approval

The investigation has received ethical approval from the institutional review board, ensuring the protection of

participants' rights and compliance with the relevant ethical guidelines.

References

[1] R. Nayak, U. C. Pati, and S. K. Das, "A comprehensive review on deep learning-based methods for video anomaly detection," *Image Vis Comput*, vol. 106, p. 104078, 2021. Elsevier. https://doi.org/10.1016/j.imavis.2020.104078.

[2] D. R. Patrikar and M. R. Parate, "Anomaly detection using edge computing in video surveillance system," *Int J Multimed Inf Retr*, vol. 11, no. 2, pp. 85–110, 2022. Springer. https://doi.org/10.1007/s13735-022-00227-8.

[3] R. Leyva, V. Sanchez, and C.-T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3463–3478, 2017. IEEE. https://doi.org/10.1109/TIP.2017.2695105.

[4] M. H. Arshad, M. Bilal, and A. Gani, "Human activity recognition: Review, taxonomy and open challenges," *Sensors*, vol. 22, no. 17, p. 6463, 2022. MDPI. https://doi.org/10.3390/s22176463.

[5] F. Harrou, N. Zerrouki, Y. Sun, and A. Houacine, "An integrated vision-based approach for efficient human fall detection in a home environment," *IEEE Access*, vol. 7, pp. 114966–114974, 2019. IEEE. https://doi.org/10.1109/ACCESS.2019.2936320.

[6] A. A. Khan *et al.*, "Crowd Anomaly Detection in Video Frames Using Fine-Tuned AlexNet Model," *Electronics (Basel)*, vol. 11, no. 19, p. 3105, 2022. MDPI. https://doi.org/10.3390/electronics11193105.

[7] K. K. Santhosh, D. P. Dogra, and P. P. Roy, "Anomaly detection in road traffic using visual surveillance: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 6, pp. 1–26, 2020. ACM Digital Library. https://doi.org/10.1145/3417989.

[8] Y. Bao, Z. Tang, H. Li, and Y. Zhang, "Computer vision and deep learning–based data anomaly detection method for structural health monitoring," *Struct Health Monit*, vol. 18, no. 2, pp. 401–421, 2019. Sage Publications. https://doi.org/10.1177/1475921718757405.

[9] Shamsi, H., Golkarieh, A., Nouri, H. *et al.* Enhanced prediction of defibrillation success in out-of-hospital cardiac arrest using nonlinear ECG features and probabilistic neural network classification. *SIViP* **19**, 647 (2025). https://doi.org/10.1007/s11760-025-04269-3

[10] S. Zaidi, B. Jagadeesh, K. V Sudheesh, and A. A. Audre, "Video anomaly detection and classification for human activity recognition," in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, IEEE, Mysore, India, 2017, pp. 544–548. https://doi.org/10.1109/CTCEEC.2017.8455012.

[11] X. Xu, J. Tang, X. Zhang, X. Liu, H. Zhang, and Y. Qiu, "Exploring techniques for vision based human activity recognition: Methods, systems, and evaluation," *sensors*, vol. 13, no. 2, pp. 1635–1650, 2013. MDPI. https://doi.org/10.3390/s130201635.

[12] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimed Tools Appl*, vol. 79, no. 41–42, pp. 30509–30555, 2020. Springer. https://doi.org/10.1007/s11042-020-09004- 3.

[13] H. Samani, C.-Y. Yang, C. Li, C.-L. Chung, and S. Li, "Anomaly detection with vision-based deep learning for epidemic prevention and control," *J Comput Des Eng*, vol. 9, no. 1, pp. 187–200, 2022. Oxford Academic. https://doi.org/10.1093/jcde/qwab075.

[14] P. Schneider, J. Rambach, B. Mirbach, and D. Stricker, "Unsupervised anomaly detection from time-of-flight depth images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 231–240.

[15] K. Doshi and Y. Yilmaz, "Fast unsupervised anomaly detection in traffic videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 624–625.

[16] B. Ramachandra, M. J. Jones, and R. R. Vatsavai, "A survey of single-scene video anomaly detection," *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 5, pp. 2293–2312, 2020. IEEE. https://doi.org/10.1109/TPAMI.2020.3040591.

[17] A. Aldayri and W. Albattah, "Taxonomy of Anomaly Detection Techniques in Crowd Scenes," *Sensors*, vol. 22, no. 16, p. 6080, 2022. MDPI. https://doi.org/10.3390/s22166080.

[18] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "Video processing using deep learning techniques: A systematic literature review," *IEEE Access*, vol. 9, pp. 139489–139507, 2021. IEEE. https://doi.org/10.1109/ACCESS.2021.3118541.

[19] J. Ren, F. Xia, Y. Liu, and I. Lee, "Deep video anomaly detection: Opportunities and challenges," in *2021 international conference on data mining workshops (ICDMW)*, IEEE, 2021, pp. 959–966. IEEE. https://doi.org/10.1109/ICDMW53433.2021.00125.

[20] R. Raja, P. C. Sharma, M. R. Mahmood, and D. K. Saini, "Analysis of anomaly detection in surveillance video: recent trends and future vision," *Multimed Tools Appl*, vol. 82, no. 8, pp. 12635–12651, 2023. Springer. https://doi.org/10.1007/s11042-022-13954-1.

[21] Sasani, F., Moghareh Dehkordi, M., Ebrahimi, Z., Dustmohammadloo, H., Bouzari, P., Ebrahimi, P., ... & Fekete-Farkas, M. (2024). Forecasting of Bitcoin Illiquidity Using High-Dimensional and Textual Features. *Computers*, *13*(1), 20.

[22] R. Wang, K. Nie, T. Wang, Y. Yang, and B. Long, "Deep learning for anomaly detection," in

Proceedings of the 13th international conference on web search and data mining, 2020, pp. 894–896. ACM Digital Library. https://doi.org/10.1145/3336191.3371876.

[23] B. R. Kiran, D. M. Thomas, and R. Parakkal, "An overview of deep learning-based methods for unsupervised and semi-supervised anomaly detection in videos," *J Imaging*, vol. 4, no. 2, p. 36, 2018. Springer. https://doi.org/10.3390/jimaging4020036.

[24] M. Baradaran and R. Bergevin, "A critical study on the recent deep learning based semi-supervised video anomaly detection methods," *Multimed Tools Appl*, pp. 1–47, 2023. Springer. https://doi.org/10.1007/s11042-023-16445-z.

[25] J. G. Munyua, G. M. Wambugu, and S. T. Njenga, "A Survey of Deep Learning Solutions for Anomaly Detection in Surveillance Videos," *International Journal of Computer and Information Technology (2279-0764)*, vol. 10, no. 5, 2021.

[26] Ahmadi, H., Mahdimahalleh, S. E., Farahat, A., & Saffari, B. (2025). Unsupervised time-series signal analysis with autoencoders and vision transformers: A review of architectures and applications. *arXiv preprint arXiv:2504.16972*.

[27] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021. ACM Digital Library. https://doi.org/10.1145/3439950.

[28] Ebrahimi, P., Dustmohammadloo, H., Kabiri, H., Bouzari, P., & Fekete-Farkas, M. (2023). Transformational entrepreneurship and digital platforms: a combination of ISM-MICMAC and unsupervised machine learning algorithms. *Big data and cognitive computing*, *7*(2), 118.

[29] Asghari, M., Shahmohamadi, P., Safaripour, A., Padash, Y., Javankiani, S., Jafarzadeh Jahromi, Z., ... & Rezaalizadeh Seresti, M. (2025). Spontaneous Pneumothorax Detection in Chest X-rays using Convolutional Neural Networks. *InfoScience Trends*, *2*(5), 71-79.

[30] Golkarieh, A., Boroujeni, S. R., Kiashemshaki, K., Deldadehasl, M., Aghayarzadeh, H., & Ramezani, A. (2025). Breakthroughs in Brain Tumor Detection: Leveraging Deep Learning and Transfer Learning for MRI-Based Classification. *Computer and Decision Making: An International Journal*, *2*, 708-722.

[31] Lashaki, R. A., Raeisi, Z., Razavi, N., Goodarzi, M., & Najafzadeh, H. (2025). Optimized classification of dental implants using convolutional neural networks and pre-trained models with preprocessed data. *BMC Oral Health*, *25*(1), 535.

[32] Jafari, M., Majidi, F., & Heydarnoori, A. (2025). Prioritizing App Reviews for Developer Responses on Google Play. *arXiv preprint arXiv:2502.01520*.

[33] F. Camarena, M. Gonzalez-Mendoza, L. Chang, and R. Cuevas-Ascencio, "An Overview of the Vision-Based Human Action Recognition Field," *Mathematical and Computational Applications*, vol. 28, no. 2, p. 61, 2023. MDPI. https://doi.org/10.3390/mca28020061.

[34] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "Anomalynet: An anomaly detection network for video surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 10, pp. 2537–2550, 2019. IEEE. https://doi.org/10.1109/TIFS.2019.2900907.

[35] W. Ullah, T. Hussain, Z. A. Khan, U. Haroon, and S. W. Baik, "Intelligent dual stream CNN and echo state network for anomaly detection," *Knowl Based Syst*, vol. 253, p. 109456, 2022. Elsevier. https://doi.org/10.1016/j.knosys.2022.109456.

[36] H. Ma and L. Zhang, "Attention-based framework for weakly supervised video anomaly detection," *J Supercomput*, pp. 1–21, 2022. Springer.

[37] H. Li, X. Shen, X. Sun, Y. Wang, C. Li, and J. Chen, "Video anomaly detection based on scene classification," *Multimed Tools Appl*, vol. 82, no. 29, pp. 45345–45365, 2023. Springer. https://doi.org/10.1007/s11042-023-15328-7.

[38] M. George, B. R. Jose, J. Mathew, and P. Kohttps://doi.org/10.1007/s11227-021-04190-9. kare, "Autoencoder-based abnormal activity detection using parallelepiped spatio-temporal region," *IET Computer Vision*, vol. 13, no. 1, pp. 23–30, 2019.

[39] V.-T. Le and Y.-G. Kim, "Attention-based residual autoencoder for video anomaly detection," *Applied Intelligence*, vol. 53, no. 3, pp. 3240–3254, 2023. Springer. https://doi.org/10.1007/s10489-022-03613- 1.

[40] L. Wang, H. Tan, F. Zhou, W. Zuo, and P. Sun, "Unsupervised anomaly video detection via a double-flow ConvLSTM variational autoencoder," *IEEE Access*, vol. 10, pp. 44278–44289, 2022. IEEE. https://doi.org/10.1109/ACCESS.2022.3165977.

[41] Y. Fan, G. Wen, D. Li, S. Qiu, M. D. Levine, and F. Xiao, "Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder," *Computer Vision and Image Understanding*, vol. 195, p. 102920, 2020. Elsevier. https://doi.org/10.1016/j.cviu.2020.102920.

[42] G. Slavic, M. Baydoun, D. Campo, L. Marcenaro, and C. Regazzoni, "Multilevel anomaly detection through variational autoencoders and bayesian models for self-aware embodied agents," *IEEE Trans Multimedia*, vol. 24, pp. 1399–1414, 2021. IEEE. https://doi.org/10.1109/TMM.2021.3065232.

[43] F. Waseem, R. P. Martinez, and C. Wu, "Visual anomaly detection in video by variational autoencoder," *arXiv preprint arXiv:2203.03872*,

2022. Cornell University. https://doi.org/10.48550/arXiv.2203.03872.

[44] Sharafkhani, F., Corns, S., & Holmes, R. (2024). Multi-step ahead water level forecasting using deep neural networks. *Water*, *16*(21), 3153.

[45] X.-G. Zhou and L.-Q. Zhang, "Abnormal event detection using recurrent neural network," in *2015 International conference on computer science and applications (CSA)*, IEEE, Wuhan, China, 2015, pp. 222–226. https://doi.org/10.1109/CSA.2015.64.

[46] C. Ma, Z. Miao, M. Li, S. Song, and M.-H. Yang, "Detecting anomalous trajectories via recurrent neural networks," in *Asian Conference on Computer Vision*, Springer, Cham, 2018, pp. 370–382. https://doi.org/10.1007/978-3-030-20870-7_23.

[47] M. Murugesan and S. Thilagamani, "Efficient anomaly detection in surveillance videos based on multi layer perception recurrent neural network," *Microprocess Microsyst*, vol. 79, p. 103303, 2020. Elsevier. https://doi.org/10.1016/j.micpro.2020.103303.

[48] M. Khalooei, M. Fakhredanesh, and M. Sabokrou, "Dominant and rare events detection and localization in video using Generative Adversarial Network," *Journal of Soft Computing and Information Technology*, vol. 8, no. 3, pp. 40–51, 2019.

[49] A. Swiecicki, N. Konz, M. Buda, and M. A. Mazurowski, "A generative adversarial network-based abnormality detection using only normal images for model training with application to digital breast tomosynthesis," *Sci Rep*, vol. 11, no. 1, p. 10276, 2021. Nature. https://doi.org/10.1038/s41598-021-89626-1.

[50] G. Pang, C. Aggarwal, C. Shen, and N. Sebe, "Editorial deep learning for anomaly detection," *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 6, pp. 2282–2286, 2022. IEEE. https://doi.org/10.1109/TNNLS.2022.3162123.

[51] W. Zhang, P. He, S. Wang, L. An, and F. Yang, "A Dynamic Convolutional Generative Adversarial Network for Video Anomaly Detection," *Arab J Sci Eng*, vol. 48, no. 2, pp. 2075–2085, 2023. Springer. https://doi.org/10.1007/s13369-022-07096-7.

[52] Boroujeni, S. R., Abedi, H., & Bush, T. (2025). Enhancing Glass Defect Detection with Diffusion Models: Addressing Imbalanced Datasets in Manufacturing Quality Control. *arXiv preprint arXiv:2505.03134*.

[53] F. Zhou, L. Wang, Z. Li, W. Zuo, and H. Tan, "Unsupervised learning approach for abnormal event detection in surveillance video by hybrid autoencoder," *Neural Process Lett*, vol. 52, pp. 961–975, 2020. Springer. https://doi.org/10.1007/s11063-019-10113-w.

[54] C.-W. Chang, C.-Y. Chang, and Y.-Y. Lin, "A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection," *Multimed Tools Appl*, vol. 81, no. 9, pp. 11825–11843, 2022. Springer. https://doi.org/10.1007/s11042-021-11887-9.

[55] L. Jie, C. Jiahao, Z. Xueqin, Z. Yue, and L. I. N. Jiajun, "One-hot encoding and convolutional neural network-based anomaly detection," *Journal of Tsinghua University (Science and Technology)*, vol. 59, no. 7, pp. 523–529, 2019.

[56] Q. Liu and X. Zhou, "A Fully Connected Network Based on Memory for Video Anomaly Detection," in *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, IEEE, Chengdu, China, 2022, pp. 221–226. https://doi.org/10.1109/CCIS57298.2022.10016377.

[57] S. Digikar, A. Chaudhari, P. Angre, and R. Pathak, "Autoencoder Based Anomaly Detection in Surveillance Videos," *Open Access International Journal of Science & Engineering (OAIJSE)*, vol. 6, pp. 29–32, 2021. DOI: 10.51397/OAIJSE06.2021.0037.

[58] K. Deepak, S. Chandrakala, and C. K. Mohan, "Residual spatiotemporal autoencoder for unsupervised video anomaly detection," *Signal Image Video Process*, vol. 15, no. 1, pp. 215–222, 2021. Springer. https://doi.org/10.1007/s11760-020-01740-1.

[59] X. He, F. Yuan, T. Liu, and Y. Zhu, "A video system based on convolutional autoencoder for drowning detection," *Neural Comput Appl*, pp. 1–13, 2023. Springer. https://doi.org/10.1007/s00521-023-08526-9.

[60] H. S. Modi and D. A. Parikh, "An intelligent unsupervised anomaly detection in videos using inception capsule auto encoder," *The Imaging Science Journal*, pp. 1–18, 2023. Taylor & Francis. https://doi.org/10.1080/13682199.2023.2202577.

[61] M. Xu, X. Yu, D. Chen, C. Wu, and Y. Jiang, "An efficient anomaly detection system for crowded scenes using variational autoencoders," *Applied Sciences*, vol. 9, no. 16, p. 3337, 2019. MDPI. https://doi.org/10.3390/app9163337.

[62] G. Slavic, A. S. Alemaw, L. Marcenaro, D. M. Gomez, and C. Regazzoni, "A Kalman Variational Autoencoder Model Assisted by Odometric Clustering for Video Frame Prediction and Anomaly Detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 415–429, 2022. IEEE. https://doi.org/10.1109/TIP.2022.3229620.

[63] X. Zhou *et al.*, "Spatial–contextual variational autoencoder with attention correction for anomaly detection in retinal OCT images," *Comput Biol Med*, vol. 152, p. 106328, 2023. Elsevier. https://doi.org/10.1016/j.compbiomed.2022.106328.

[64] Ackerson, J. M., Dave, R., & Seliya, N. (2021). Applications of Recurrent Neural Network for Biometric Authentication & Anomaly

Detection. *Information*, *12*(7), 272. https://doi.org/10.3390/info12070272

[65] M. Yagan, E. A. Yilmaz, and H. Özkan, "Anomaly Detection in Surveillance Videos Using Regression with Recurrent Neural Networks," in *2022 30th Signal Processing and Communications Applications Conference (SIU)*, IEEE, Safranbolu, Turkey, 2022, pp. 1–4. https://doi.org/10.1109/SIU55565.2022.9864893.

[66] N. Elsayed, Z. ElSayed, and A. S. Maida, "LiteLSTM Architecture Based on Weights Sharing for Recurrent Neural Networks," *arXiv preprint arXiv:2301.04794*, 2023. Taylor & Francis.https://doi.org/10.1080/1206212X.2025.2499869.

[67] T. Ganokratanaa and S. Aramvith, "Generative adversarial network for video anomaly detection," in *Generative Adversarial Networks for Image-to-Image Translation*, Elsevier, 2021, pp. 377–420. https://doi.org/10.1016/B978-0-12-823519-5.00011-7.

[68] T. N. Nguyen and J. Meunier, "Hybrid deep network for anomaly detection," *arXiv preprint arXiv:1908.06347*, 2019. Cornell University. https://doi.org/10.48550/arXiv.1908.06347.

[69] Z. Liu, Y. Nie, C. Long, Q. Zhang, and G. Li, "A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13588–13597.

[70] A. Alia, M. Maree, and M. Chraibi, "A hybrid deep learning and visualization framework for pushing behavior detection in pedestrian dynamics," *Sensors*, vol. 22, no. 11, p. 4040, 2022. MDPI. https://doi.org/10.3390/s2211404.0

[71] Alafif, T., Hadi, A., Allahyani, M., Alzahrani, B., Alhothali, A., Alotaibi, R., & Barnawi, A. (2023). Hybrid Classifiers for Spatio-Temporal Abnormal Behavior Detection, Tracking, and Recognition in Massive Hajj Crowds. *Electronics*, *12*(5), 1165. https://doi.org/10.3390/electronics12051165

[72] M. Sabokrou, M. Fathy, and M. Hoseini, "Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder," *Electron Lett*, vol. 52, no. 13, pp. 1122–1124, 2016. Wiley Online Library. https://doi.org/10.1049/el.2016.0440.

[73] Y. Li, Y. Cai, J. Liu, S. Lang, and X. Zhang, "Spatio-temporal unity networking for video anomaly detection," *IEEE Access*, vol. 7, pp. 172425–172432, 2019. IEEE. https://doi.org/10.1109/ACCESS.2019.2954540.

[74] B. Ramachandra and M. Jones, "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*, 2020, pp. 2569–2578.

[75] X. Wang *et al.*, "Robust unsupervised video anomaly detection by multipath frame prediction," *IEEE Trans Neural Netw Learn Syst*, vol. 33, no. 6, pp. 2301–2312, 2021. IEEE. https://doi.org/10.1109/TNNLS.2021.3083152.

[76] S. Chang, Y. Li, S. Shen, J. Feng, and Z. Zhou, "Contrastive attention for video anomaly detection," *IEEE Trans Multimedia*, vol. 24, pp. 4067–4076, 2021. IEEE. https://doi.org/10.1109/TMM.2021.3112814.

[77] S. Biswas and R. V. Babu, "Real time anomaly detection in H. 264 compressed videos," in *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, IEEE, Jodhpur, India, 2013, pp. 1–4. https://doi.org/10.1109/NCVPRIPG.2013.6776164.

[78] K. Xu, T. Sun, and X. Jiang, "Video anomaly detection and localization based on an adaptive intra-frame classification network," *IEEE Trans Multimedia*, vol. 22, no. 2, pp. 394–406, 2019. IEEE. https://doi.org/10.1109/TMM.2019.2929931.

[79] Y. Zhang, H. Lu, L. Zhang, X. Ruan, and S. Sakai, "Video anomaly detection based on locality sensitive hashing filters," *Pattern Recognit*, vol. 59, pp. 302–311, 2016. Elsevier. https://doi.org/10.1016/j.patcog.2015.11.018.

[80] Y. Hao, J. Li, N. Wang, X. Wang, and X. Gao, "Spatiotemporal consistency-enhanced network for video anomaly detection," *Pattern Recognit*, vol. 121, p. 108232, 2022. Elsevier. https://doi.org/10.1016/j.patcog.2021.108232.

[81] R. J. Franklin and V. Dabbagol, "Anomaly detection in videos for video surveillance applications using neural networks," in *2020 Fourth International Conference on Inventive Systems and Control (ICISC)*, IEEE, Coimbatore, India, 2020, pp. 632–637. https://doi.org/10.1109/ICISC47916.2020.9171212.

[82] S. Parameswaran, J. Harguess, C. Barngrover, S. Shafer, and M. Reese, "Evaluation schemes for video and image anomaly detection algorithms," in *Automatic Target Recognition XXVI*, SPIE, 2016, pp. 98–109. https://doi.org/10.1117/12.2224667.

[83] Raeisi, Z., Sodagartojgi, A., Sharafkhani, F., Roshanzamir, A., Najafzadeh, H., Bashiri, O., & Golkarieh, A. (2025). Enhanced classification of tinnitus patients using EEG microstates and deep learning techniques. *Scientific Reports*, *15*(1), 15959.