Review

Detection of Abnormal Human Behavior Using Unsupervised Learning in Video Surveillance Systems

Naejoung Kwak , Seongsoo Cho * , Cheolhee Yoon *

Posted Date: 8 January 2025

doi: 10.20944/preprints202501.0576.v1

Keywords: Unsupervised learning; abnormal behavior detection; video surveillance; generative adversarial networks (GANs)

Review

Detection of Abnormal Human Behavior Using Unsupervised Learning in Video Surveillance Systems

Naejoung Kwak [1], Seongsoo Cho [2,*] and Cheolhee Yoon [3,*]

[1] Department of Information Security, Pai Chai University, Daejeon, 35345, Korea
[2] Department of Applied Mathematics, Kongju National University, Gongju 32588, Korea
[3] Laboratory of Autonomous Vehicle and Block-chain, Korean National Police University, Chungnamm, 31539, Republic of Korea
* Correspondence: css3617@gmail.com (S.C.); bertter@police.ac.kr (C.Y.)

Abstract: As public safety demands increase, there is a growing need for intelligent surveillance systems capable of detecting abnormal human behavior in real-world settings. Numerous detection techniques based on machine learning and deep learning models have been developed for abnormal behavior detection. Among these, unsupervised learning enables anomaly detection without labeled data by learning normal patterns and identifying deviations, making it a practical solution to address the shortage of abnormal data. This paper surveys and analyzes recent unsupervised learning techniques for detecting abnormal human behavior in surveillance video streams, reviewing commonly used datasets, discussing practical limitations, and identifying areas for improvement to enhance the reliability and efficiency of unsupervised models in surveillance applications.

Keywords: unsupervised learning; abnormal behavior detection; video surveillance; generative adversarial networks (GANs)

1. Introduction

The development of IT technology has provided convenience in life, but has also increased the need for security. Accordingly, interest in research on methods to detect and prevent abnormal behaviors using surveillance camera systems for human safety in public places such as streets, parks, and subway stations has increased.

In the recognition of abnormal behaviors in videos acquired from surveillance cameras, human behavior rcognition is performed by analyzing spatial and visual information and extracting features [1,2]. Traditional behavior recognition methods extract features only at important feature points where behaviors occur frequently, and apply the extracted features to various pattern classifiers to perform behavior recognition. Machine learning algorithms such as Bayesian networks [3], support vector machines (SVMs) [4], and random forests (RFs) [5] have shown good performance as pattern classifiers. However, machine learning algorithms have problems such as low accuracy because their performance is heavily dependent on the extracted features and they cannot cope with various changes in objects or behaviors [6].

Recently, research on applying deep learning algorithms to behavior recognition methods has been active. Deep learning is a multi-stage learning process that automatically extracts, recognizes, and classifies representative features from multiple hidden layers [7]. Recently, with the increase in computing resources and available data, deep learning has been applied to the fields of behavior recognition and abnormal behavior recognition. It has also shown very efficient performance in video surveillance systems [8].

 doi:10.20944/preprints202501.0576.v1

Deep learning techniques for detecting abnormal human behavior can be classified into three methods: supervised learning, partially supervised learning, and unsupervised learning [9]. In supervised learning, the model analyzes and learns normal and abnormal behavior patterns from the input layer temporally and spatially during the training phase [10–12]. The models detect behavior by identifying deviations from learned patterns or comparing new data with identified behavior clusters. Partially supervised learning is a method of learning using a partially labeled dataset, and can be classified into weakly supervised learning and semi-supervised learning [13]. Semi-supervised learning can create a model using labeled data and then classify unlabeled data using the model to create labels. Unsupervised learning is a method of training a model using unlabeled data, and since there is no correct answer for the input data, the model learns the patterns of the data on its own. Unsupervised learning is used as an alternative to solve the problem of lack of abnormal behavior data in the field of abnormal behavior detection along with semi-supervised learning because it can detect abnormal behavior even without data labels [14]. Unsupervised learning methods include reconstruction-based methods [15] and generative-based methods [16]. Among the unsupervised learning methods, the reconstruction-based method analyzes only normal event data and detects abnormalities by checking for low reconstruction errors [17]. The generative-based method artificially generates images using trained distribution patterns, and the model effectively solves the problem of lack of data by distinguishing whether the images are real or fake [18]. However, both the reconstruction-based and generative approaches have difficulty identifying specific abnormal behaviors and are very sensitive to environmental changes [19].

There have been several previous studies investigating abnormal human behavior detection systems. Verma et al. [20] investigated the detection of abnormal behaviors in single objects and crowds. They summarize supervised and unsupervised learning methodologies based on various feature extraction algorithms and SVM, HMM, and ANN classifiers. However, this paper lacks recent papers on methods using deep learning. Patrikar et al. [21] provide a survey on abnormal detection systems in video surveillance and edge computing-based abnormal detection. This paper also divides the study into two parts: learning and modeling. However, this paper lacks exploration of methods using unsupervised learning and focuses on the application of edge computing. Shubber et al. [22] detects abnormal behaviors by dividing them into machine learning and deep learning methods. This paper considers fights and assaults as abnormal behaviors. It presents data sets for each method and compares the performance of each method. However, this paper does not mention unsupervised learning using deep learning. Also, it does not cover other abnormal behaviors besides violence. Roka et al. [23] present a recent study on anomalous behavior detection. The authors categorize anomalous behaviors into statistical-based, data mining-based, and machine learning-based techniques, and provide a description of the various techniques for anomalous behavior detection, along with their pros and cons. However, the paper lacks a description of what behaviors exist in the reference dataset. Choudhry et al. [24] also comprehensively describe machine learning methods, dividing them into three major categories: supervised learning, semi-supervised learning, and unsupervised learning. The paper presents the challenges of detecting anomalous behaviors using machine learning techniques in the future. However, the scope is too broad and does not focus on image-based detection. Altowairqi et al. [25] investigate Crowd Anomaly Detection. The paper also presents a comparison of different types of crowd detection that match the dataset used. It also explains that unsupervised learning methods such as AE and GAN can detect anomalous behaviors that have not been learned in advance. However, the discussion of unsupervised methods lacks a description of how to detect unseen data and the methods for doing so. Tay et al.[26] investigate existing methods for abnormal behavior detection and deep learning approaches for sensor-based and vision-based inputs. This paper focuses on abnormal behaviors such as falls that can occur in daily life. It also investigates available public datasets and provides solutions to the problem of lack of datasets for abnormal behaviors that occur in daily life. This paper is limited to abnormal behaviors in daily life and lacks investigations on a wider range of abnormal behavior detection. Jahan and Islam[27] provide a critical review of video-based human activity recognition (HAR) techniques, followed by

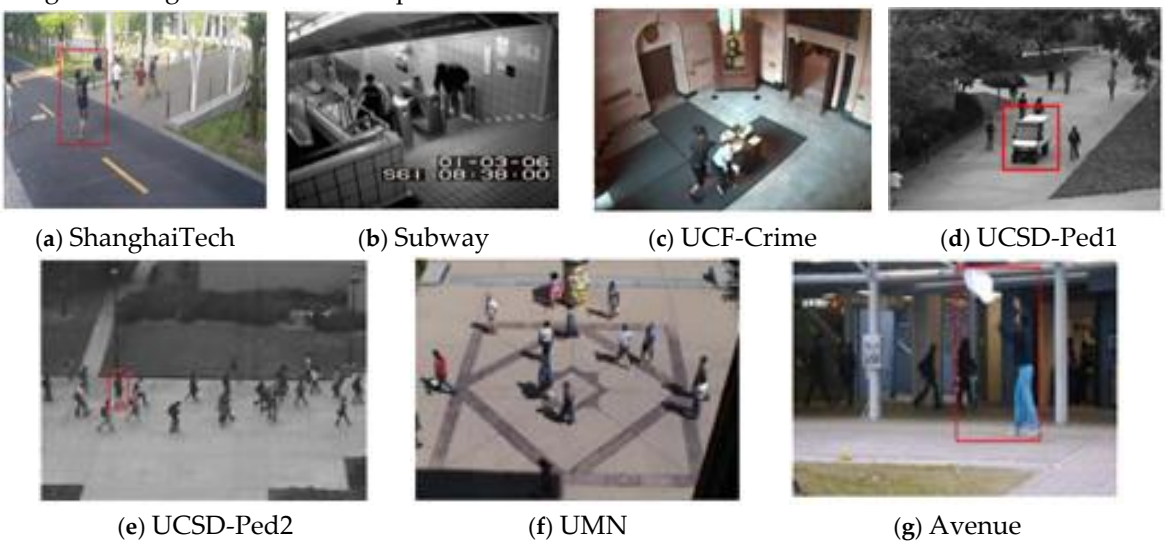
 doi:10.20944/preprints202501.0576.v1

an analysis of machine learning and deep learning techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), hidden Markov models (HMMs), and K-means clustering. This paper investigates and explains unsupervised learning in machine learning, but does not explain unsupervised learning methods in deep learning.

These results from the analysis of related studies show that there is still much unexplored research. In addition, there is not much investigation on abnormal behavior recognition methods that apply unsupervised learning. In this paper, we examine the results of a study on unsupervised learning techniques applied to the automation of abnormal behavior detection in surveillance cameras. We also examine open datasets used to train the model. We also examine open research issues in the application of unsupervised learning in the field of abnormal human behavior detection in surveillance cameras.

2. Dataset

Several datasets have been used to benchmark research methods for human abnormal behavior recognition. Figure 1 shows a sample of each dataset.



(**a**) ShanghaiTech (**b**) Subway (**c**) UCF-Crime (**d**) UCSD-Ped1

(**e**) UCSD-Ped2 (**f**) UMN (**g**) Avenue

Figure 1. Samples of abnormal behaviors in each data set.

The ShanghaiTech (ST) Campus dataset [28] was created by ShanghaiTech Univesity in 2017 and consists of 13 scenes with complex lighting conditions and camera angles from all sides. The dataset consists of 330 training videos with only normal events and 107 test videos with 130 abnormal events. The dataset has a total of 317,398 frames and was collected using an RGB camera with a resolution of 856 × 480 at 24 FPS overlooking a pedestrian walkway.

Subway dataset [29] was introduced by Adam et al. in 2008. It consists of two videos, totaling 2 hours, containing 209 and 150 frames, which are exit gate videos and entrance gate videos, respectively. The videos are recorded in grayscale format at 15 FPS with a resolution of 512 × 384, and there are a total of 125,475 frames. This dataset contains 19 types of unusual events, such as walking in the wrong direction, wandering, no payment, people jumping or cutting in at turnstiles, and janitors cleaning the walls.

The UCF-Crime dataset [30] was created by the Computer Vision Department at the University of Central Florida. This dataset consists of 1900 videos, 128 hours, of 1900 Internet videos captured by multiple RGB cameras at various locations. The abnormal behaviors include 13 real-world abnormalities, including abuse, arrest, arson, assault, traffic accident, theft, explosion, fight, robbery, shooting, theft, and vandalism. This dataset can be used for two tasks: event recognition of the 13 group activities and detection of abnormal behaviors in each specific group.

The UCSD Anomaly Detection Dataset [31] consists of 70 video footages acquired from above to monitor pedestrian walkways. This dataset was created at the University of California, San Diego in

 doi:10.20944/preprints202501.0576.v1

2013. This dataset consists of two video sets, called Ped1 and Ped2, which are grayscale image sequences recorded at 10FPS. Ped1 is created at a resolution of 38×158, and Ped2 is created at a resolution of 360×240. Ped1 contains scenes of people walking toward the camera and moving away from the camera, various perspective distortions, cyclists, skaters, etc. Ped2 contains 16 training videos and 12 test videos, which contain 12 abnormal events. It contains scenes of pedestrians moving parallel to the camera plane from an upper angle, and non-human objects detected are considered abnormal.

The UMN dataset [32] was created by R. Mehran et al. in 2009 at the University of Minnesota and consists of 11 different abnormal event scenarios, including 3 indoor and 3 outdoor scenes. All videos have the same frame rate of 30 FPS and were recorded at a resolution of 640 × 480 using a static camera. This dataset contains a total of 22 videos for training and testing, consisting of 7739 frames.

The Avenue dataset [33] was created by the Chinese University of Hong Kong (CUHK) in 2013 and contains a total of 37 videos, including 16 training video clips and 21 test video clips. The resolution of each image sequence is 640 × 360 and the frame rate is 25 FPS. The footage consists of 30,652 frames, evenly split between training and testing, and features 14 unique events, including people running, wandering, and throwing objects.

3. Unsupervised Learning-Based Abnormal Behavior Detection Method

Deep learning is an artificial intelligence (AI) method that models the behavior of the human brain and implements data processing on a computer. Deep learning models can recognize complex patterns of images, text, sounds, and other data and predict the results. Deep learning techniques are used to automate tasks that generally require human intelligence, such as describing images or converting audio files into text. Among deep learning techniques, unsupervised learning methods learn patterns, structures, and unique features of data without labels specified for the data. Unsupervised learning models are widely used in three main tasks: clustering, association, and dimensionality reduction. Unsupervised learning is also applied to human behavior recognition to distinguish between normal and abnormal behaviors. In abnormal behavior detection, unsupervised learning can be used when obtaining labels for various abnormal behaviors is difficult or costly [34,35]. Unsupervised learning methods can be divided into reconstruction-based and generative-based approaches. Reconstruction-based methods learn patterns from input images, while generative-based methods attempt to generate artificial images based on learned patterns. Table 1 shows some recent unsupervised learning methods.

3.1. Reconstruction-Based Methods

Reconstruction-based methods train the model using only normal data, modeling the distribution of this data. Abnormal data is assigned a high reconstruction error by the model. At the inference stage, if the test image is abnormal, the model encounters difficulty in reconstructing it. Reconstruction-based detection methods include autoencoders (AE) and variational autoencoders (VAE).

3.1.1. Autoencoder

An autoencoder is a neural network that learns input data and attempts to reconstruct new images based on previously learned patterns. An AE consists of two components: an encoder and a decoder. The objective of this model is to minimize the reconstruction error, enabling it to reconstruct images more accurately based on the learned data. Figure 2 illustrates the structure of an AE.

 doi:10.20944/preprints202501.0576.v1

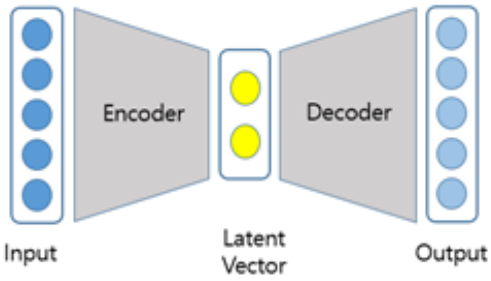


Figure 2. Structure of an Autoencoder.

Hasan et al. [36] utilized a sparse coding autoencoder to preserve spatiotemporal information between input and output. They employed a 2D convolutional network to encode sequences of grayscale 2D images from each segment of a video, using single-channel input images stacked in the temporal dimension. Medel et al. [37] applied a reconstruction-based method for anomaly detection, enhancing spatiotemporal information with a convolutional long short-term memory (Conv-LSTM) network. Sabokrou et al. [38] used two types of autoencoders: a regular autoencoder and a sparse autoencoder, which constrains dimensional features to retain the most useful active neurons in the latent layer. Zhao et al. [39] employed a 3D convolutional autoencoder to preserve temporal information and track spatial features across the temporal dimension, also incorporating data augmentation to increase the number of training samples. Zhou et al. [40] proposed SC2NET, a novel network for feature learning that combines both motion and appearance features of images. This network computes sparsity loss and learns to construct useful spatiotemporal features by learning from configuration error.

Recently, Wang et al. [45] and Sampath et al. [46] proposed spatiotemporal AEs, achieving AUC values above 0.98 for detecting abnormal behaviors on the UCSD Ped1 and Ped2 datasets. However, spatiotemporal AE cannot fully utilize and understand implicit video information, especially when using a single modality camera. To address this limitation, Liu et al. [47] proposed an anomaly detection method using an object-centric scene inference network. They identify abnormal behaviors based on the assumption that autoencoders yield high configuration error scores for abnormal instances. However, this assumption does not always hold, and reconstruction error scores are sometimes lower than expected. Gong et al. [48] addressed this issue by treating each encoding feature as a query to the decoder network, storing all normal encoding features in memory. The decoder retrieves the closest normal encoding in memory for each query instance, resulting in high reconstruction errors for abnormal instances that cannot map to the closest normal encoding.

Autoencoders are used to extract features from data, but the large dataset size poses challenges in obtaining a representative distribution of normal samples. Liu et al. [49] proposed a method that applies fully connected layers of convolutional neural networks to memory modules. Autoencoders learn and reconstruct feature representations from datasets, with a large number of loss functions input to the memory module for scoring to identify abnormal images. This approach requires well-tuned score thresholds for various environments to effectively classify behavior categories. Yan et al. [50] implemented a memory clustering autoencoder to detect abnormal human behavior. The autoencoder reconstructs the input sequence, while the clustering and scoring systems distinguish abnormal human behaviors in video. However, this method faces the challenge of training conflicts between the autoencoder and clustering components.

3.1.2. Variational Autoencoder

Variational autoencoders (VAEs) function as probabilistic generative models, requiring a neural network composed of an encoder and a decoder. The encoder's primary role is to adjust the parameters of the variational distribution, while the decoder maps from the latent space to the input space. VAEs are a key component in probabilistic graphical models and variational Bayesian methods [51]. Figure 3 illustrates the structure of a VAE.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2025 doi:10.20944/preprints202501.0576.v1

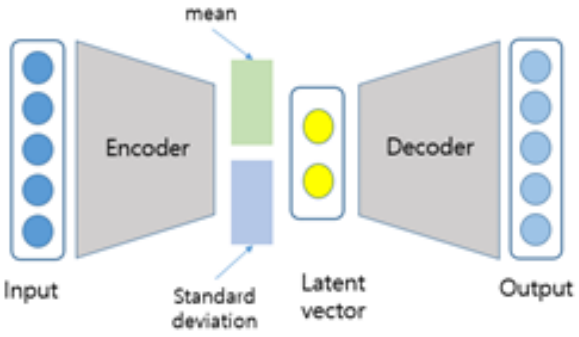


Figure 3. Structure of a Variational Autoencoder.

Wang et al. [52] proposed a memory-based autoencoder that differentiates abnormal images by exploiting reconstruction errors based on memory. They developed a cognitive memory augmentation network (CMAN) to implement memory-based recognition and judgment.

Wang et al. [53] proposed a generative neural network using two VAEs to detect abnormal behavior in complex scenes. The first VAE quickly filters normal samples at the input layer, and the second VAE extracts hierarchical features, fusing low-level and high-level samples. This approach achieves a near-perfect AUC of 0.999 in certain scenarios. However, detecting abnormal events in video sequences remains challenging due to the complexity of frame data. However, it is very difficult to detect abnormal events in video sequences due to the complexity of frame data.

To address this, Yan et al. [54] proposed a two-stream structure consisting of appearance and motion streams. The VAE calculates anomaly scores based on reconstruction error probability and generates latent variables via sampling. This method achieves an AUC of 0.913, though it requires additional computational time for processing images.

Wang et al. [55] introduced a double-flow convolutional long short-term memory VAE to predict normal video sequences in an unsupervised, distributed manner. This VAE calculates the average reconstruction probability for abnormal behavior detection, achieving an AUC of 0.888 by combining long-term and short-term memory methods. However, this model has difficulty in predicting very small foreground target objects.

Cho et al. [56] developed an implicit two-path autoencoder with distribution modeling of normal features based on a regularization flow model for AHB detection in an unsupervised manner. Using the Ped2 dataset, this approach achieves an AUC score of 0.992, and on the CUHK dataset, it achieves 0.880, demonstrating high performance. However, due to the similarity in pedestrian shapes, motions, and walking patterns, distinguishing between normal and abnormal scenes remains difficult for VAEs and regularization flow models.

Liu et al. [57] proposed a probabilistic video normality network to learn various normal event patterns across temporal, spatial, and spatiotemporal dimensions. This model encodes past frames into a posterior distribution, sampling latent variables with a VAE to predict future frames. This approach achieves an AUC of 0.984 on the Ped2 dataset and 0.907 on the CUHK dataset. However, its performance depends on the optimal hyperparameter settings for AHB detection.

Table 1. Comparison of reconstruction-based method.

Ref.	Year	Method	Performance (AUC)		
			Ped1	Ped2	Avenue
[36]	2016	Autoencoder	0.81	0.9	0.7
[41]	2019	Autoencoder	0.897	0.913	N/A
[42]	2022	Autoencoder	N/A	0.956	N/A
[43]	2021	Autoencoder	N/A	0.972	0.879

 doi:10.20944/preprints202501.0576.v1

[44]	2022	Autoencoder	0.907	0.977	0.894
[45]	2023	Autoencoder	N/A	0.984	0.861
[46]	2023	Autoencoder	0.902	0.997	N/A
[47]	2023	Autoencoder	N/A	0.983	0.917
[49]	2022	Autoencoder	N/A	0.968	0.875
[50]	2023	Autoencoder	0.907	0.977	0.894
[52]	2021	variational Autoencoder	N/A	0.962	N/A
[54]	2020	variational Autoencoder	0.75	0.91	0.79
[55]	2022	variational Autoencoder	0.884	0.888	0.872
[56]	2022	variational Autoencoder	N/A	0.992	0.880
[57]	2023	variational Autoencoder	N/A	0.984	0.907
[58]	2022	variational Autoencoder	N/A	N/A	0.862

3.1.3. Summary

In reconstruction-based methods, autoencoders are commonly used for image dimensionality reduction. This learning process then computes a loss function within the network, which helps identify abnormal behaviors in the input data. The autoencoder model's effectiveness depends on the distribution and quality of the data. In addition, in an environment with many objects, it is difficult for autoencoders to reconstruct each object and predict abnormal behaviors

To address the generalization issues arising from data distribution, methods such as regularization and probabilistic formulation have been incorporated into the VAE approach. These adaptations enable VAEs to detect abnormal behaviors in more heterogeneous environments. However, VAE has difficulty in reconstructing small objects due to the probabilistic calculation across the entire image. Models that combine certain spatiotemporal features can detect abnormal behaviors effectively, though they tend to perform worse than supervised learning methods, especially with long-distance activities and complex scenes. In addition, all types of models have problems such as high computational cost, the need for hyperparameter tuning, and difficulty in generalizing to diverse data sets.

3.2. Generative-Based Methods

3.2.1. GAN

Generative-based methods attempt to generate artificial images based on patterns learned by the model. A neural network called a generative adversarial network (GAN) [59] is commonly used in this approach. A GAN is a generative model consisting of two main components: a generator and a discriminator, as illustrated in Figure 4. The generator creates new instances based on the statistical properties of the training data, while the discriminator determines whether the input is from the generator (fake) or the training data (real). In anomal human behavior recognition, the difference between the original and generated images is used to detect abnormal human behavior in the captured frame. Since this approach does not require labeled data, it falls under the category of unsupervised learning. Figure 4 illustrates the structure of a GAN

 doi:10.20944/preprints202501.0576.v1

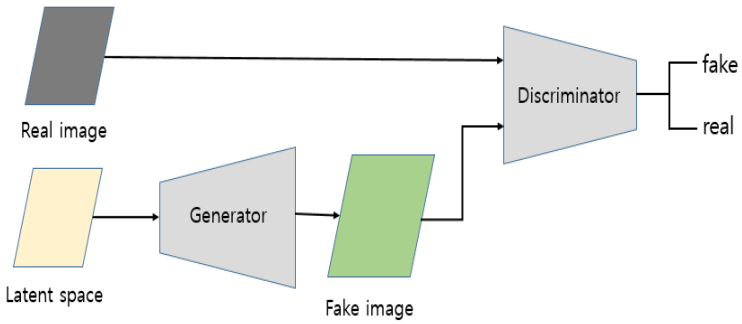


Figure 4. Structure of a GAN.

Patil et al. [60] proposed Moving Object Segmentation (MOS) to detect moving individuals. This model uses two generator adversarial networks (RMS-GAN) to perform repetitive segmentation of moving objects and estimate foreground human figures. GAN estimates the probability map for foreground human objects, achieving an AUC score of up to 0.95. However, this model has difficulty in estimating accurate motion compared to the foreground.

Yang et al. [61] introduced a bidirectional retrospective generative adversarial network (BR-GAN) for abnormal behavior detection. In this model, the generator performs bidirectional prediction, including both forward and backward predictions, enhancing long-term motion consistency estimation for human objects across frames. This approach achieved an AUC of up to 0.976. However, since the initial frame of the video must be specified as the initial input, this model cannot detect abnormal behaviors in the first few seconds. In addition, if the input image is distorted, the model has difficulty in processing small human objects.

Yu et al. [18] proposed adversarial event prediction (AEP) for detecting rare patterns in input behaviors. AEP combines reconstruction-based and generative-based detection methods, as reconstruction-based methods often struggle with diverse training data. Adversarial networks are employed to provide predictions for various environmental settings in the dataset. GAN is used to predict past and future frames, achieving an AUC of 0.979 in HAB recognition. However, AEP has limitations due to its inability to recognize the background during preprocessing.

Zhao et al. [62] introduced a block-level background reference frame (BRF) to address overlapping backgrounds, along with a foreground reference frame (FRF) to infer human objects. In this model, GAN predicts the current location of the foreground object after preprocessing the BRF in the input image. However, in some scenarios, the integration of BRF and FRF is lacking, increasing computational complexity.

Ganokratanaa et al. [63] proposed a GAN-based method for spatiotemporal anomaly detection and localization in surveillance videos, achieving an AUC score of 0.996 on the UMN dataset, demonstrating near-perfect performance. However, this method has difficulty in detecting anomalies when the abnormal events and normal patterns are similar.

Aslam et al. [64] proposed a two-stream, attention-based approach capable of end-to-end learning. During inference, they compute a normality score using a reconstruction-based method, and in the prediction phase, they use GAN to enhance feature learning performance. This approach achieved AUC scores of 0.869 on the ST dataset and 0.894 on the AVENUE dataset, representing the best performance among studies using these datasets.

Table 2. Comparison of Generative-based methods.

Ref.	Year	Performance (AUC)		
		Ped1	Ped2	Avenue
[18]	2022	0.979	0.979	0.949

 doi:10.20944/preprints202501.0576.v1

[61]	2021	0.847	0.976	0.886
[63]	2022	0.988	0.976	0.908
[64]	2022	0.907	0.977	0.894
[65]	2022	N/A	0.963	0.871
[66]	2022	0.981	0.801	0.735
[67]	2021	0..892	0.892	N/A
[68]	2022	0.969	0.969	0.866
[69]	2022	0.975	0.971	0.947
[70]	2023	N/A	0.977	0.897
[71]	2023	0.921	0.976	0.897
[72]	2023	N/A	0.968	0.887

3.2.2. Summary

The generative detection approach excels at recognizing new environments that were not present during training, making it suitable for detecting previously untrained abnormal events. It can also enhance the quality of training data and images, minimizing overlap between foreground and background. However, since the model contains both a generator and discriminator, computational complexity is increased. Additionally, a large dataset is required for training, and small objects, due to their limited pixel representation, provide less information for learning, potentially reducing the model's generalization capability.

4. Discussion

When applying unsupervised learning to detect abnormal behavior in camera-based surveillance systems, certain challenges must be addressed along with the advantages.

4.1. Data

Data issues are a primary challenge in unsupervised learning for detecting abnormal behavior. One of the biggest hurdles is the data distribution, which varies depending on the environment where detection takes place. Environments such as open spaces, crowded areas, or closed rooms can each have distinct patterns, meaning that the model must adapt its understanding to specific environments. Reconstruction-based methods, in particular, rely on learned weights based on these initial conditions, making them sensitive to changes in data distribution.

Accurately detecting abnormal behavior within a limited timeframe can also be problematic, as models may struggle to identify anomalies within the constraints of individual frames. This is especially true for methods like VAEs, which often have limitations in pixel location accuracy. To improve results, research is focused on developing enhanced techniques for detecting abnormal behaviors accurately and quickly.

As human behavior patterns vary widely, especially in abnormal situations, frequent changes can make it difficult for the model to recognize established patterns. Adding more data to accommodate these changes can increase dataset size, but too much data may reduce the model’s ability to generalize from patterns, leading to overfitting. This calls for strategies that balance dataset expansion and generalization to ensure reliable anomaly detection across different scenarios and environments.

4.2. Occlusion

 doi:10.20944/preprints202501.0576.v1

Occlusion is a significant challenge in computer vision, especially for surveillance in crowded spaces. When multiple objects or individuals overlap, it becomes increasingly difficult to discern specific behaviors or identify if an action is abnormal. Occlusion makes it challenging to isolate and analyze individual actions, especially when the behavior happens in dense crowds.

Unsupervised learning methods add complexity to this challenge, as they only learn from the patterns of input images. Consequently, the model may learn crowd patterns but fail to recognize individual actions within the crowd that deviate from the norm. Distinguishing abnormal behavior amidst occlusion requires innovative approaches, such as multi-view or depth-aware analysis, which can help disambiguate objects and improve anomaly detection in crowded environments. Additionally, improvements in tracking technologies that leverage optical flow or motion consistency could assist the model in focusing on individual movements, even when occlusion occurs.

To advance in this area, researchers are exploring new models that consider crowd density, position tracking, and object separation to ensure accurate detection of anomalous human behavior within crowded or occluded scenes.

4.3. Small Target Detection

When detecting abnormal behavior in small targets, the reduced pixel representation presents a unique challenge. Smaller objects provide limited visual information, making it difficult for the model to understand key details necessary for detecting anomalies. This lack of detail leads to two primary issues. First, it can be difficult for the model to determine if the target is a human object or something else entirely. Second, when there is limited information about the human target's features, the model struggles to understand the nature of any detected behavior, whether abnormal or not. This is a problem for both reconstruction-based and generation-based methods, as both rely on sufficient data representation to identify patterns and anomalies.

The challenge of small target detection requires solutions that maximize the model's sensitivity to minimal information while preserving accuracy. Approaches such as super-resolution techniques, which artificially increase the resolution of small objects, could help make these details more recognizable to the model. Other potential solutions include implementing attention mechanisms that prioritize areas with small, high-priority objects, allowing the model to focus on features essential for detecting anomalies.

Research is also exploring lightweight networks with heightened sensitivity to detect small object movements and actions. Addressing this issue is essential for effective abnormal behavior detection in complex environments where targets are small, such as those observed at long distances in surveillance footage.

4.4. High Computational Resources

Unsupervised learning models for anomaly detection typically employ two types of approaches: reconstruction-based methods, which consist of an encoder and decoder, and generation-based methods, which consist of a generator and discriminator. The structure of these models can lead to substantial computational requirements, especially as model complexity increases to improve detection accuracy. This complexity demands high computational power, making it less feasible for real-time surveillance or deployment in resource-limited environments.

To address this, research is exploring ways to reduce model complexity without sacrificing performance. For example, an integrated model that combines the generator and discriminator into a single efficient framework could streamline the anomaly detection process. Optimizing network architectures and using model compression techniques, such as pruning or quantization, are also promising approaches to reduce computational load.

Additionally, lightweight models are gaining attention due to their efficiency advantages. Lightweight models are optimized for faster inference, quicker decision-making, and reduced computational requirements, making them more suitable for real-time applications. Lightweight HAB detection models can help address the challenges of high resource demand, allowing for more

 doi:10.20944/preprints202501.0576.v1

scalable and efficient anomaly detection in various surveillance scenarios. Further research in this area aims to develop adaptive models that balance detection accuracy with computational efficiency, especially for deployment in edge devices or in distributed surveillance networks.

5. Conclusions

In this paper, we investigated and reviewed state-of-the-art methods for abnormal human behavior detection using unsupervised learning techniques, categorizing them into reconstruction-based and generative-based approaches. Reconstruction-based methods, such as autoencoder and variational autoencoder, focus on identifying high reconstruction errors in abnormal data, leveraging the patterns learned from normal data. These approaches have shown promise in applications with well-structured data but face challenges when dealing with complex environments, small target detection, and high computational demands. Generative-based methods, represented by generative adversarial networks (GANs), provised a robust alternative by simulating normal and abnormal scenarios. These methods have demonstrated impressive adaptability in detecting anomalies in unseen environments but require substantial computational resources and large datasets for effective performance. We also reviewed the most commonly used datasets for training and benchmarking abnormal behavior detection models, including ShanghaiTech, Subway, UCF-Crime, UCSD Ped1 and Ped2, UMN, and Avenue datasets.

Based on this, several open research issues on abnormal human behavior detection were discussed. One prominent issue is ensuring model robustness across diverse environments while addressing overfitting, as adapting to varied data distributions remains a significant concern. Additionally, occlusion in crowded scenes poses substantial difficulties, as overlapping objects hinder the accurate analysis of individual behaviors, necessitating the development of advanced techniques such as multi-view or depth-aware approaches. Another ongoing challenge is the detection of small targets, where models struggle to capture and analyze fine details due to limited visual information. Finally, the high computational demands of current models, particularly those based on generative adversarial networks (GANs), underscore the need for more lightweight and computationally efficient architectures to enable broader applicability and real-time processing.

In conclusion, while unsupervised learning methods have significantly advanced the field of abnormal behavior detection in video surveillance systems, they still face limitations that need to be addressed. Future research should focus on developing robust, scalable models that balance accuracy with computational efficiency. Incorporating innovative approaches such as multi-modal learning, self-supervised learning, and edge-based deployment could help overcome existing challenges. Furthermore, expanding and diversifying public datasets to better reflect real-world conditions will be crucial in enhancing model performance and generalization. By addressing these challenges, the field can move closer to realizing highly reliable, real-time surveillance systems capable of ensuring public safety in diverse environments.

Author Contributions: Conceptualization, N.K. and S.C.; methodology, N.K.; software, N.K. and S.C.; validation, N.K., S.C. and C.Y.; formal analysis, N.K.; investigation, N.K.; resources, S.C.; data curation, S.C.; writing—original draft preparation, N.K.; writing—review and editing, S.C.; visualization, S.C.; supervision, Y.L.; project administration, C.Y.; funding acquisition, C.Y. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Data recorded in the current study are available in all tables and figures of the manuscript.

Acknowledgments: This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00337489, Development of data drift management technology to overcome performance degradation of AI analysis models).

 doi:10.20944/preprints202501.0576.v1

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kim, D., Kim, H., Mok, Y., and Paik, J. (2021). Real-time surveillance system for analyzing abnormal behavior of pedestrians. *Applied Sciences*, *11*(13), 6153.
2. Patwal, A., Diwakar, M., Tripathi, V., and Singh, P. (2023). An investigation of videos for abnormal behavior detection. *Procedia Computer Science*, *218*, 2264-2272.
3. Xiao, Q., and Song, R. (2018). Action recognition based on hierarchical dynamic Bayesian network. *Multimedia Tools and Applications*, *77*, 6955-6968.
4. Abidine, B. M. H., Fergani, L., Fergani, B., and Oussalah, M. (2018). The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition. *Pattern Analysis and Applications*, *21*(1), 119-138.
5. Hu, C., Chen, Y., Hu, L., and Peng, X. (2018). A novel random forests based class incremental learning method for activity recognition. *Pattern Recognition*, *78*, 277-290.
6. Hu, X., Hu, S., Huang, Y., Zhang, H., and Wu, H. (2016). Video anomaly detection using deep incremental slow feature analysis network. *IET Computer Vision*, *10*(4), 258-267.
7. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90.
8. Oyedotun, O. K., and Khashman, A. (2017). Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*, *28*(12), 3941-3951.
9. Pandiaraja, P., Saarumathi, R., Parashakthi, M., and Logapriya, R. (2023, March). An analysis of abnormal event detection and person identification from surveillance cameras using motion vectors with deep learning. In *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1225-1232). IEEE.
10. Ahn, J., Park, J., Lee, S. S., Lee, K. H., Do, H., and Ko, J. (2023). SafeFac: Video-based smart safety monitoring for preventing industrial work accidents. *Expert Systems with Applications*, *215*, 119397.
11. Onyema, E. M., Balasubaramanian, S., Iwendi, C., Prasad, B. S., and Edeh, C. D. (2023). Remote monitoring system using slow-fast deep convolution neural network model for identifying anti-social activities in surveillance applications. *Measurement: Sensors*, *27*, 100718.
12. Ullah, H., and Munir, A. (2023). Human activity recognition using cascaded dual attention cnn and bi-directional gru framework. *Journal of Imaging*, *9*(7), 130.
13. Ren, J., Xia, F., Liu, Y., and Lee, I. (2021, December). Deep video anomaly detection: Opportunities and challenges. In *2021 international conference on data mining workshops (ICDMW)* (pp. 959-966). IEEE.
14. Zhang, C., Li, G., Xu, Q., Zhang, X., Su, L., and Huang, Q. (2022). Weakly supervised anomaly detection in videos considering the openness of events. *IEEE transactions on intelligent transportation systems*, *23*(11), 21687-21699.
15. Wang, Y., Qin, C., Bai, Y., Xu, Y., Ma, X., and Fu, Y. (2022, November). Making reconstruction-based method great again for video anomaly detection. In *2022 IEEE International Conference on Data Mining (ICDM)* (pp. 1215-1220). IEEE.
16. Ganokratanaa, T., Aramvith, S., and Sebe, N. (2019, November). Anomaly event detection using generative adversarial network for surveillance videos. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1395-1399). IEEE.
17. Nayak, R., Pati, U. C., and Das, S. K. (2021). A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing*, *106*, 104078.
18. Yu, J., Lee, Y., Yow, K. C., Jeon, M., and Pedrycz, W. (2022). Abnormal event detection and localization via adversarial event prediction. *IEEE transactions on neural networks and learning systems*, *33*(8), 3572-3586.
19. Şengönül, E., Samet, R., Abu Al-Haija, Q., Alqahtani, A., Alturki, B., and Alsulami, A. A. (2023). An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey. Applied Sciences, 13(8), 4956.
20. Verma, K.K., Singh, B.M. and Dixit, A. (2022). A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. Inf. tecnol. 14, 397–410.
21. Patrikar , D. R., and Parate, M. R. (2022). Anomaly detection using edge computing in video surveillance system. Multimedia Information Retrieval, 11(2), 85–110.
22. Shubber, M. S. M., and Al-Ta'i, Z. T. M. (2022). A review on video violence detection approaches. Nonlinear Analysis and Applications, 13, 1117–1130.
23. Roka, S., Diwakar, M., Singh, P., and Singh, P. (2023). Anomaly behavior detection analysis in video surveillance: a critical review. Electronic Imaging, 32(4), 042106.1-042106.21
24. Choudhry, N., Abawajy, J., Huda, S., and Rao, I. (2023). A comprehensive survey of machine learning methods for surveillance videos anomaly detection. IEEE Access, 11, 114680-114713.
25. Altowairqi, S., Luo, S., and Greer, P. (2023). A review of the recent progress on crowd anomaly detection. Advanced Computer Science and Applications, 14(4), 659-669.

 doi:10.20944/preprints202501.0576.v1

26. Tay , N. C., Connie , T., Ong , T. S., Teoh, A. B. J., and The. (2023). P. S. A Review of Abnormal Behavior Detection in Activities of Daily Living. IEEE Access, 11, 5069-5088.
27. Jahan, S., and Islam, M. R. (2024). A Critical Analysis on Machine Learning Techniques for Video-based Human Activity Recognition of Surveillance Systems: A Review. arXiv preprint arXiv:2409.00731.
28. Luo, W., Liu, W., and Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision* (pp. 341-349).
29. Adam, A., Rivlin, E., Shimshoni, I., and Reinitz, D. (2008). Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, *30*(3), 555-560.
30. Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6479-6488).
31. Li, W., Mahadevan, V., and Vasconcelos, N. (2013). Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, *36*(1), 18-32.
32. Mehran, R., Oyama, A., and Shah, M. (2009, June). Abnormal crowd behavior detection using social force model. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 935-942). IEEE.
33. Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision* (pp. 2720-2727).
34. Cinelli, L. P., Marins, M. A., Da Silva, E. A. B., and Netto, S. L. (2021). *Variational methods for machine learning with applications to deep networks* (Vol. 15). Springer.
35. Jebur, S. A., Hussein, K. A., Hoomod, H. K., Alzubaidi, L., and Santamaría, J. (2022). Review on deep learning approaches for anomaly event detection in video surveillance. *Electronics*, *12*(1), 29.
36. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. (2016). Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 733-742).
37. Medel, J. R., and Savakis, A. (2016). Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390*.
38. Sabokrou, M., Fathy, M., and Hoseini, M. (2016). Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, *52*(13), 1122-1124.
39. Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., and Hua, X. S. (2017, October). Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 1933-1941).
40. Zhou, J. T., Di, K., Du, J., Peng, X., Yang, H., Pan, S. J., Tsang, I. W., Liu, Y., Qin, Z., and Goh, R. S. M. (2018, April). Sc2net: Sparse lstms for sparse coding. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).
41. Wang, T., Miao, Z., Chen, Y., Zhou, Y., Shan, G., and Snoussi, H. (2019). AED-Net: An Abnormal Event Detection Network. Engineering, 5, 930–939
42. Wang, B., and . Yang , C. (2022). Video Anomaly Detection Based on Convolutional Recurrent AutoEncoder. Sensors, 22(12), 4647.
43. Tang, W.; Feng, Y.; Li, J. (2021). An autoencoder with a memory module for video anomaly detection. In *Proceedings of the 2021 36th Youth Academic Annual Conference of Chinese Association of Automation (* pp. 473–478).
44. Aslam, N.; Rai, P.K.; Kolekar, M.H. (2022). A3N: Attention-based adversarial autoencoder network for detecting anomalies in video sequence. Vis. Commun. Image Represent. 87, 103598.
45. Wang, Y., Liu, T., Zhou, J., and Guan, J. (2023). Video anomaly detection based on spatio-temporal relationships among objects. *Neurocomputing*, *532*, 141-151.
46. Sampath, D. K., and Kumar, K. (2023). Abnormal Crowd Behaviour Detection in Surveillance Videos Using Spatiotemporal Inter-Fused Autoencoder. *International Journal of Intelligent Engineering & Systems*, *16*(6).
47. Liu, Y., Guo, Z., Liu, J., Li, C., and Song, L. (2023). Osin: Object-centric scene inference network for unsupervised video anomaly detection. *IEEE Signal Processing Letters*, *30*, 359-363.
48. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Hengel, A. V. D. (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* (pp. 1705-1714).
49. Liu, Q., and Zhou, X. (2022, November). A Fully Connected Network Based on Memory for Video Anomaly Detection. In *2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS)* (pp. 221-226). IEEE.
50. Yan, M., Xiong, Y., and She, J. (2023). Memory clustering autoencoder method for human action anomaly detection on surveillance camera video. *IEEE Sensors Journal*, *23*(18), 20715-20728.
51. Cinelli, L. P., Marins, M. A., Da Silva, E. A. B., and Netto, S. L. (2021). *Variational methods for machine learning with applications to deep networks* (Vol. 15). Springer.
52. Wang, T., Xu, X., Shen, F., and Yang, Y. (2021). A cognitive memory-augmented network for visual anomaly detection. *IEEE/CAA Journal of Automatica Sinica*, *8*(7), 1296-1307

 doi:10.20944/preprints202501.0576.v1

53. Wang, T., Qiao, M., Lin, Z., Li, C., Snoussi, H., Liu, Z., and Choi, C. (2018). Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security*, *14*(5), 1390-1399.
54. Yan, S., Smith, J. S., Lu, W., and Zhang, B. (2020). Abnormal event detection from videos using a two-stream recurrent variational autoencoder. *IEEE Transactions on Cognitive and Developmental Systems*, *12*(1), 30-42.
55. Wang, L., Tan, H., Zhou, F., Zuo, W., and Sun, P. (2022). Unsupervised anomaly video detection via a double-flow convlstm variational autoencoder. *IEEE Access*, *10*, 44278-44289.
56. Cho, M., Kim, T., Kim, W. J., Cho, S., and Lee, S. (2022). Unsupervised video anomaly detection via normalizing flows with implicit latent features. *Pattern Recognition*, *129*, 108703.
57. Liu, Y., Yang, D., Fang, G., Wang, Y., Wei, D., Zhao, M., Cheng, K., Liu, J., and Song, L. (2023). Stochastic video normality network for abnormal event detection in surveillance videos. Knowledge-Based Systems, 280, 110986.
58. Slavic, G., Baydoun, M., Campo, D., Marcenaro, L., Regazzoni, C. (2022). Multilevel Anomaly Detection Through Variational Autoencoders and Bayesian Models for Self-Aware Embodied Agents. IEEE Trans. Multimed. 24, 1399–1414.
59. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.
60. Patil, P. W., Dudhane, A., and Murala, S. (2020). End-to-end recurrent generative adversarial network for traffic and surveillance applications. *IEEE Transactions on Vehicular Technology*, *69*(12), 14550-14562.
61. Yang, Z., Liu, J., and Wu, P. (2021). Bidirectional retrospective generation adversarial network for anomaly detection in videos. *IEEE Access*, *9*, 107842-107857.
62. Zhao, L., Wang, S., Wang, S., Ye, Y., Ma, S., and Gao, W. (2021). Enhanced surveillance video compression with dual reference frames generation. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1592-1606.
63. Ganokratanaa, T., Aramvith, S., and Sebe, N. (2022). Video anomaly detection using deep residual-spatiotemporal translation network. *Pattern Recognition Letters*, *155*, 143-150.
64. Aslam, N., Rai, P. K., and Kolekar, M. H. (2022). A3N: Attention-based adversarial autoencoder network for detecting anomalies in video sequence. *Journal of Visual Communication and Image Representation*, *87*, 103598.
65. Li, D., Nie, X., Li, X.;,Zhang, Y., Yin, Y. (2022). Context-related video anomaly detection via generative adversarial network. Pattern Recognit. Lett, 156, 183–189.
66. Zhang, Z., Zhong, S.h., Fares, A., Liu, Y. (2022). Detecting abnormality with separated foreground and background: Mutual generative adversarial networks for video abnormal event detection. Comput. Vis. Image Underst, 219, 103416.
67. Alafif, T., Alzahrani , Cao, B., Alotaibi, Y. R., Barnawi, A., and Chen, M. Generative adversarial network based abnormal behavior detection in massive crowd videos: a Hajj case study. Journal of Ambient Intelligence and Humanized Computing, 13(8), 4077-4088.
68. Hao , Y., Li, J., Wang, N., Wang, X., and Gao, X. (2022). Spatiotemporal consistency-enhanced network for video anomaly detection. Pattern Recognition, 121, 108232.
69. Yu, J., Kim, J.-G.;, Gwak, J., Lee, B.-G., Jeon, M. (2022). Abnormal event detection using adversarial predictive coding for motion and appearance. Inf. Sci, 586, 59–73.
70. Huang, H.; Zhao, B.; Gao, F.; Chen, P.; Wang, J.; Hussain, A. (2023). A Novel Unsupervised Video Anomaly Detection Framework Based on Optical Flow Reconstruction and Erased Frame Prediction. Sensors, 23, 4828
71. Huang, C.; Wen, J.; Xu, Y.; Jiang, Q.; Yang, J.; Wang, Y.; Zhang, D. (2023). Self-Supervised Attentive Generative Adversarial Networks for Video Anomaly Detection. IEEE Trans. Neural Netw. Learn. Syst., 34, 9389–9403.
72. Li, G.; He, P.; Li, H.; Zhang, F. (2023). Adversarial composite prediction of normal video dynamics for anomaly detection. Comput. Vis. Image Underst. 232, 103686