Journal of Computer Science and Electrical Engineering
ISSN: 2663-1946
DOI: 10.61784/jcsee3006

THEORY AND APPLICATIONS OF VIDEO ABNORMAL BEHAVIOR DETECTION

PeiChen Wu[1], DengBin Xu[1], LiNing Yuan[2*]
[1]*School of Information Network Security, People's Public Security University of China, Beijing 100038, China.*
[2]*School of Public Security Big Data Modern Industry, Guangxi Police College, Nanning 530028, Guangxi, China.*
Corresponding Author: LiNing Yuan, Email: yuanlining@gxjcxy.edu.cn

Abstract: Video abnormal behavior detection is a research hotspot in the field of computer vision. By extracting the spatiotemporal characteristics of video content, we can determine whether there are abnormal events and their types in the video, and identify the location and time of the abnormal events. Based on supervised/unsupervised learning, this paper systematically combs and summarizes the existing video abnormal behavior detection methods. Starting from the current mainstream modeling idea, the supervision method is described in detail, and the completely unsupervised method is introduced to train the model. The network architectures of different models are compared, and the characteristics of various anomaly detection models in terms of test data sets, usage scenarios, advantages and limitations are summarized. Then, through common evaluation criteria such as frame level standard and pixel level standard, the model is compared and the performance is evaluated. At the same time, the performance of different methods is compared within the class, and the results are analyzed and summarized in depth. Finally, the future development direction is outlined briefly, and the development trend of video anomaly detection from virtual composite dataset, multi-modal large-scale model to lightweight model is discussed.
Keywords: Abnormal behavior detection; Deep learning; Fully unsupervised; Multimodal features

1 INTRODUCTION

Video surveillance is the process of analyzing and processing video data collected by surveillance to detect abnormal behavior in the video [1]. Video abnormal behavior detection refers to the analysis and processing of video data collected by surveillance to detect abnormal behavior or events in the video. Detecting abnormal behavior in surveillance video by hand is tedious and laborious, and the massive surveillance data cannot be detected one by one by manpower [2]. With the lower cost of deployment, video surveillance is now widely used in many scenarios, such as preventing public hazards, guarding key locations, and fighting crime. The intelligent and accurate processing of surveillance data has become a direction of concern for researchers. In recent years, video anomalous behavior detection techniques have shown excellent performance driven by deep learning [3] but there is still room for improvement in terms of accuracy, reliability and scalability [4].

Early machine learning algorithms learned shallow features from video data and achieved significant results in video anomalous behavior detection through e.g. normal modeling. Random Forest [5], Bayesian networks [6], Markov models [7] and support vector machines [8] etc. are used to understand and recognize target behaviors, these methods rely on preprocessing and hand-crafted features, require a lot of time and resources to process, and do not scale well to different datasets, and show poor performance in practical applications [9]. Compared with traditional machine learning methods, deep learning is a multi-stage learning process. On the one hand, multiple hidden layers are used to automatically extract representative features for a specific task, and on the other hand, normal behavior is learned through features to determine abnormalities. Supervised methods are applied through autoencoder [10], Generative Adversarial Network (GAN) [11] and VIT (Vision Transformer) [12]. Extracting features from the end-to-end framework and calculating deviations from the normal model to discriminate whether it is anomalous or not. Fully unsupervised methods, as a strict form of unsupervised methods, model normal events using Gaussian distribution as opposed to previous unsupervised methods, and further by training binary classifiers [13] and generating pseudo-labels [14] and other methods to discriminate abnormal events.

This paper focuses on abnormal behavior detection methods, commonly used datasets, evaluation criteria, and experimental comparisons. There are four contributions as follows: (1) From unsupervised and completely unsupervised, it systematically analyzes the recent progress of video abnormal behavior detection methods. (2) The commonly used datasets and evaluation criteria are introduced in detail from the perspective of practical applications. (3) Comparison and performance evaluation of the main methods are carried out. (4) Analyze and summarize the future direction of current video abnormal behavior detection algorithms.

2 OVERVIEW OF VIDEO ABNORMAL BEHAVIOR DETECTION

Video anomalies are defined based on real scenarios and scenes, and are usually categorized into two types of anomalies: normal actions that occur at restricted locations in a scene, and abnormal actions that occur at arbitrary locations in a scene. This means that an activity that is abnormal in one scene becomes normal in another scene, for

example, cycling along a bike path is normal, while cycling on a sidewalk becomes abnormal behavior. Video anomalous behavior detection techniques are those that use limited a priori knowledge [15] and hidden patterns or structures in the data to analyze and process the video data to identify abnormal events or behaviors that do not conform to the normal behavioral pattern [16] in order to improve the efficiency and accuracy of the surveillance system and detect potential security risks or abnormalities in a timely manner. Abnormal behavior detection can be regarded as the process of discriminating outliers, and the discriminating general formula is:

$$F(y)=\begin{cases}\text{Nomal} & D(F(y),P_N)\le\tau\\ \text{Abnormal} & D(F(y),P_N)>\tau\end{cases} \tag{1}$$

where D represents the distance metric, PN represents the process of detecting as anomalous, the normal distribution of normal behavioral data obtained through training [10], F represents a feature extractor mapping raw video data to a set of discriminative features, and τ represents the threshold.

Due to the high dimensionality and diversity of the data, the process of simulating the fit PN and calculating the distance metric D through F construction is a critical step in anomalous behavior detection. Early anomalous behavior detection methods for F [17] were based on manual features for classification. Nowadays, they are gradually developed to learn deep features for abnormal behavior detection through autoencoders, convolutional networks and classifiers, etc. [18-20].

For PN and D, earlier approaches based on fully convolutional neural networks [20] evaluated the Mahalanobis distance by Gaussian estimation of the feature maps on the training data and distinguished anomalies using an assumption of an approximate Gaussian distribution. In recent years, deep neural networks (DNNS) such as generative adversarial networks and coder-decoders are mostly used for implicit learning PN *and* D [17]. autoencoder based methods [21] use generative adversarial networks to model normal distributions of normal behavioral data and act as anomaly discriminators to evaluate the distance metric D. Some researchers use coder-decoder networks to detect out-of-distribution data [22] and thus derive the distance metric D.

3 CLASSIFICATION OF ABNORMAL BEHAVIOR DETECTION METHODS

In this paper, we categorize the abnormal behavior detection algorithms in terms of both supervised / unsupervised learning methods. In supervised methods, the algorithms are categorized by distance-based methods, probability-based methods and reconstruction-based methods. In the fully unsupervised approach, classification is done by pseudo-label generation and training of classifiers as a differentiation from the traditional unsupervised approach. Figure 1 summarizes the classification of abnormal behavior detection methods.

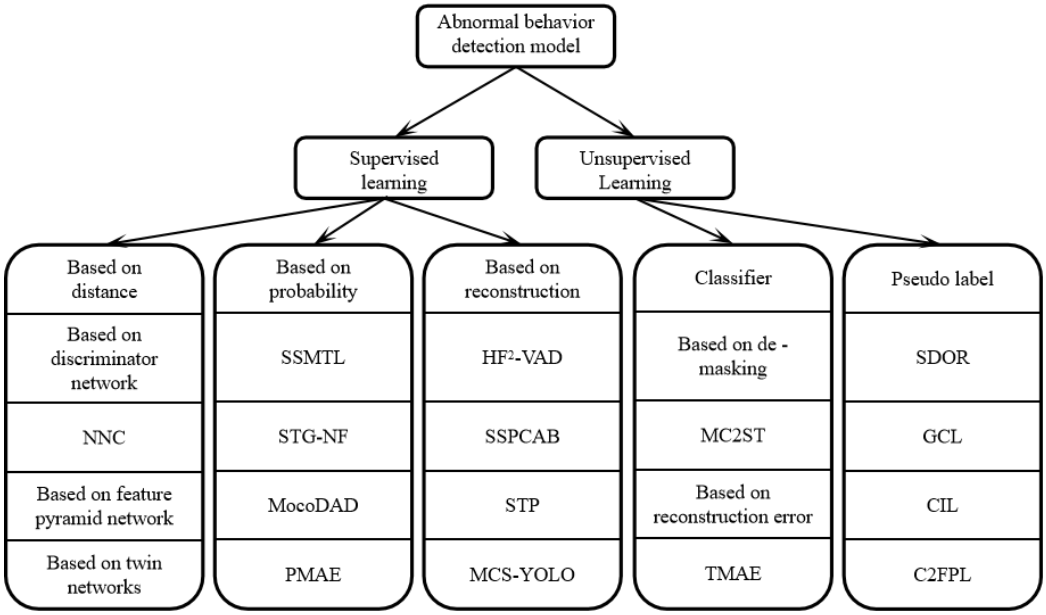

Figure 1 Summary of Classification of Anomaly Detection Methods

3.1 Supervision Available

According to different modeling approaches the deviation calculation mean methods are further classified into distance based methods, probability based methods, and reconstruction based methods. Traditional methods earlier relied on manual features such as foreground occlusion,flow histograms [24], magnitude of motion [25], Histogram of Gradients (HOG) [26], Dense Trajectories [27] and Space Time Interest Point (STIP) [28]. These features rely on a priori knowledge and have poor descriptive capabilities. With the rise of deep learning in computer vision tasks [29, 30] in computer vision tasks, there has been a tendency in recent years to extract features from end-to-end frameworks, including autoencoders, generative adversarial networks [31] and Vision Transformer (VIT) [32].

The distance-based approach determines the anomaly score by creating a normality model using the training data and measuring the deviation from the model, i.e., the outliers. The distance-based approach uses many different types of features, including manual features and features extracted by deep learning. Many different measures of outliers are also used, with earlier approaches typically using a hybrid Gaussian function to model normal feature vectors, and then using the Mahalanobis distance to compute the outliers. In recent years single class SVMs are used to compute decision boundaries for normal training video feature vectors.The disadvantage of this method is that updating the model given new training data is costly and requires the SVM algorithm to be re-computed on all old and new data. The distance based methods in recent years are analyzed in general through Table 1.

Table 1 Overall Analysis of Distance-based Methods

arithmetic	Test Data Set	Applicable Scenarios	dominance	limitations
Sabokrou et al.[33]	UCSDPed2	Open Outdoor	Strong model generalization	Generators require a high level of training
NNC [34]	CUHK Avenue Subway	Open outdoor/indoor	Fast detection speed	Poor detection of samples when the light varies
Ionescu et al. [35]	UCSD Ped2 CUHKAvenue Shanghaitech	Open Outdoor	Fast detection; high accuracy	Poor model generalization
Ramachandra et al.[36]	UCSDPed1 UCSDPed2 CUHK Avenue	Open Outdoor	Distance function generalizes well as a plug-and-play module; good model robustness	The training process of the model has dependency on the annotated data

The following summarizes the distance-based approaches in recent years. The discriminator network based approach [33] in which the use of adversarial training for video anomalous behavior detection is proposed. This is done by reconstructing a discriminator network to distinguish between the original image sequences and the noise sequences, which are obtained by a denoising autoencoder, which also acts as a generator in it. Since the autoencoder is trained only on image sequences from the training data, outliers can be extracted so that the discriminator network can distinguish between anomalous image sequences. Narrowed Normality Clusters (NNC) [34] is a two-stage anomaly detection framework. In the feature extraction phase, first, NNC extracts fixed-size spatio-temporal features from the training video and augments the spatio-temporal features with appearance features extracted by Convolutional Neural Networks (CNN). In the first stage of detection, a robust normal representation is created by performing K-means clustering to find clusters representing different types of normal motion and appearance features and eliminating small clusters corresponding to outliers. In the second stage of detection, the method trains a one-class SVM on each cluster to narrow down the boundaries of the remaining clusters, and test samples are recognized as abnormal if they fail to satisfy the maximum normality score of any of the one-class SVMs. Based on Feature Pyramid Network (FPN) [37] method [35] Based on the previous[34] proposed to convert the video abnormal behavior detection problem into a binary classification problem. Firstly, features are extracted by FPN, convolutional selfencoder is trained and K-means clustering is performed based on image feature information and backpropagation, and then multiple single-class vector machines are trained[38] as one-to-one binary classifiers. If the classifier of the test sample is less than the highest score of the other classifiers it is an anomaly. A twin network based approach [36] On the basis of simple nearest neighbor model [23] The twin network is trained by training the twin network and replacing the handcrafted features and distance functions with the features and distance functions learned by the twin network. The twin network is used to classify pairs of video sequences as same or different and if the video sequence of the test sample is not similar to any normal video sequence, it is recognized as anomalous. Probabilistic-based methods are computed in a certain probability space with outliers from a model, such as Probabilistic Graphical Models (PGMS) or Gaussian mixtures of probability distributions[39]. Most probabilistic methods rely on the application of traditional features in traditional models, such as spatio-temporal gradients[40], optical flow fields[41] and STIP with Markov random fields[42] and Gaussian mixture models[43] in combination. There are some methods that use deep learning and have obtained high accuracy, these methods have both advantages such as robustness and generalization, but also suffer from slow detection speed. A general analysis of probability-based methods in recent years is presented through Table 2.

Table 2 Overall analysis of probability-based methods

arithmetic	Test Data Set	Applicable Scenarios	dominance	limitations
SSMTL [44]	CUHK Avenue Shanghaitech UBnormal	Open outdoor/street	Simple network structure	The model performs poorly during multi-task training
STG-NF [45]	Shanghaitech UBnormal	Open outdoor/indoor	Algorithm generalization	Model targets skeletal computational anomalies and is poor at detecting non-anthropogenic anomalous behavior
PMAE [47]	UCSD Ped1	Open outdoor/street/indoor	Strong model generalization and good robustness	Low detection accuracy for some scenarios

The following summarizes the probability-based approaches in recent years. Self-supervised Multi-task Learning (SSMTL) [44] is a video anomalous behavior detection framework based on a multi-head self-attentive module.SSMTL first obtains the object bounding box through a target detector to create an object-centric time series. Then, the model is trained by joint self-supervised learning of four agent subtasks, and multi-head self-attention modules are introduced to process input information at different times. Finally, the different subtasks give anomaly scores according to a predefined strategy in the prediction phase and the average of all the anomaly scores is used as the final anomaly score. Spatio-Temporal Graph Normali-zing Flows (STG-NF) [45] is an anomaly detection method based on normalized flows of human poses.STG-NF firstly performs pose estimation and tracking on the input video, and represents each person as a temporal pose graph by extracting the key points of the human body, and then processes each pose sequence composed of temporal pose graphs.The model learns a bi-directional mapping of the data distribution to the latent Gaussian distribution during training, and finally estimates the probability of each pose sequence during inference to determine the probability of each pose sequence [48]. of the bi-directional mapping, and finally the probability of each pose sequence is estimated during inference to obtain the anomaly score. Motion Conditioned Diffusion Anomaly Detection (MocoDAD) [46] MocoDAD is a video anomaly detection model based on denoising diffusion probabilistic models, which assumes that both normal and anomalies are multi-species.MocoDAD predicts multi-species human motion trajectories by considering the human body's keypoints and utilizing denoising diffusion probabilistic models (DDPMs) to generate different but reasonable human motion trajectories using the generalization ability of the model's diffusion process [49] to predict multi-category human motion trajectories, and using the generalization ability of the model diffusion process to generate different but reasonable motion trajectories.The model, after statistically aggregating the predicted motion trajectories, determines an abnormality when the actual motion does not match with the set of predicted motion trajectories. Probabilistic Memory Self-encoding (PMAE) [47] A semi-supervised abnormal behavior detection network based on probabilistic memory model is designed to solve the problem of extreme imbalance between normal and abnormal behavior data.PMAE uses causal three-dimensional convolution and temporal dimension shared fully connected layer in extracting spatio-temporal features in order to avoid confusion of predicted information and ensure the temporal order of information. In terms of auxiliary modules, the quality of video frame reconstruction in the backbone network is improved by probabilistic model and memory module, and the gap between predicted future frames and real frames is utilized to measure the degree of abnormality. Reconstruction-based methods use features learned from normal videos as inputs, and reconstruct the inputs using a representation of the learned features when testing the samples, thus comparing the differences between the reconstruction results and the test samples to determine whether there is an anomaly or not [50]. The premise of this method is that reconstructing inputs outside the normal distribution based on feature representations learned from normal is very difficult, so it is reasonable to use reconstruction error as an anomaly score [51-53]. Most reconstruction-based methods employ deep learning, such as convolutional autoencoders [54-56] and generative adversarial networks [57-59]. This approach also has drawbacks: first, the model has to be retrained to adapt if the training set is updated, and second, most models do not evaluate the spatial localization of the anomaly [60] of the anomalies. Reconstruction-based methods in recent years are analyzed in general through Table 3.

Table 3 Overall Analysis of Reconstruction-based Methods

Arithmetic	Test Data Set	Applicable Scenarios	dominance	limitations
EVAL [61]	CUHKAvenue Shanghaitech	Open outdoor/street	High model accuracy and generalizability	Inaccurate categorization of subjects of aberrant behaviour
HF2-VAD [63]	UCSDPed2 CUHKAvenue	Open Outdoor	Novel integration strategy; high detection accuracy	Poor detection accuracy for anomalous objects at long distances
SSPCAB [64]	CUHKAvenue Shanghaitech	Open Outdoors	High module portability	For some methods the effect is not obvious
STP [65]	UCSDPed2 CUHKAvenue Shanghaitech	Open outdoor/street	The model has significantly improved the spatial localization of anomalous behaviors	Changes in the proportion of feature inputs have an effect on detection effectiveness

The following summarizes reconstruction-based approaches in recent years. Explainable Video Anomaly Localization (EVAL) is a single-scene video anomaly localization framework [61]. EVAL is a single-scene video anomaly localization framework, which first constructs a standard model of a new scene by deep learning a general representation of objects and their motions, and forming an example for the new scene by calculating the appearance, direction of motion, speed, and background score. Finally, for videos of the same scene, it compares the computed features with the standard model of the scene to determine whether it is anomalous or not. Hybrid framework that integrates Flow reconstruction and Frame prediction seamlessly to handle Video Anomaly Detection, HF2-VAD [63] Flow reconstruction and frame prediction are integrated to handle Video Anomaly Detection. Firstly, Multi-Level Memory modules in an Autoencoder with Skip Connections,ML MemAE SC, are designed to store normal patterns for optical flow reconstruction, so that anomalous events with large flow reconstruction errors can be sensitively recognized. events. HF2 -VAD uses a Conditional Variational Auto Encoder (CVAE) in the context of stream

reconstruction to predict the next frame for a given number of frames and to capture correlations between video frames and the optical flow. In CVAE, the quality of the stream reconstruction inherently affects the quality of the frame prediction, so reconstructing poorly anomalous event optical streams further affects the quality of future frames and thus detects anomalies. Self-supervised Predictive Convolutional Attentive Block (SSPCAB) [64] Integrates reconstruction-based functionality with powerful portability for easy incorporation into a variety of state-of-the-art anomaly detection methods.SSPCAB learns to detect masked central regions in the sample through a convolutional layer with an expansion filter, and the generated activation maps are passed through the channel attention block. In addition, SSPCAB designs a loss function that minimizes the reconstruction error in detecting the masked region of the sample. Spatio-temporal Predictive (STP) [65] It is a method for predicting anomaly detection based on spatio-temporal normality.STP first performs spatio-temporal feature extraction on the input video according to the original resolution, after which it is cropped and passed to the encoder to learn the potential representations of the features, and then inputs four decoding branches to further predict the spatio-temporal features, and takes the error between the predicted features and the actual features as the normality. Finally, the reconstruction error between the comparison and the normal samples is used as the basis for judging the anomalies in the testing stage.

3.2 Completely Unsupervised

With the explosive growth of video data and the practical application of video anomaly behavior detection, marking all anomalies in a video gradually becomes impossible. To solve this problem, Fully-Unsupervised Video Anomaly Detection (FVAD) based on completely unsupervised video anomalous behavior has attracted the interest of researchers [66]. Compared with the traditional unsupervised [67] compared to traditional unsupervised, it is characterized by the use of target variables without any labeling in the training phase, and the model can only use the features of the input data itself for learning and pattern recognition without any manually labeled supervisory information. Completely unsupervised methods in recent years are analyzed in general through Table 4.

Table 4 Overall analysis of fully unsupervised methods

arithmetic	Test Data Set	Applicable Scenarios	dominance	limitations
Tudor et al. [66]	UCSD Ped1 UCSD Ped2 CUHK Avenue Subway	Indoor/Outdoor	Detection is fast.	There is a significant gap in detection performance compared to some of the supervised models.
MC2ST [68]	UCSD Ped1 UCSD Ped2 CUHK Avenue Subway	Indoor/Outdoor	A new frame-level motion feature with better representation and generalization is proposed	There is a possibility of causing noise or getting abnormal frames during sampling.
Li et al. [69]	UCF-Crime	Indoor/Outdoor/Day/Night	Classifier training by normal behavioral clustering with detection performance comparable to semi-supervised models	Autoencoders do not perform well with some complex datasets
TMAE [70]	UCSD Ped1 UCSD Ped2 CUHK Avenue Shanghaitech	outdoors	Predicting anomalous frames by spatio-temporal features	Not effective for covert abnormal behavior
SODR [71]	UCSD Ped1 UCSD Ped2 Subway	Indoor/open outdoor	Aligning feature inputs and anomaly scoring through an end-to-end training approach	Relatively single type of detection of abnormal behavior, lack of motor characteristics application
GCL [72]	Shanghaitech	Open Outdoor	A novel approach to negative learning was adopted	The modeling system may treat rare normal events as abnormal, thus creating bias
CIL [73]	UCSD Ped1 UCSD Ped2 Subway	Indoor/Outdoor	Resolving confusion bias in pseudolabel generation; focusing on temporal relationships of features	More computational costs in generating high-quality counterfeit labels
C2FPL [14]	XD-Violence	Indoor/Outdoor/Day/Night	The two phases of labels generated complement each other	The test data set is small and more experiments are needed to further validate the performance

A de-masking based approach [66] The text author authentication technique is applied to anomaly detection. This is done by first finding a number of frames sequentially using a sliding window algorithm and assuming that the first half is normal and the second half is anomalous, then extracting visual and motion features from these frames, then removing the most discriminative features by training a classifier and iterating this step, and finally calculating anomaly scores at the end of each iteration to adjust the window accuracy, and taking the anomaly scores obtained from the last anomalous frame window The average value is calculated as the final result. Multiple Classifier Two Sample Tests

(MC2ST) [68] The link between multiple classifiers and test samples is described and a past frame sampling method is proposed to improve its testing capability.MC2ST utilizes a de-masking based method to sample from past frames and merge them with the first half of the current window to form normal frames [66]. MC2ST utilizes a sliding window to sample from past frames and merges them with the first half of the current window to form normal frames, and then extracts features to train the classifier. In addition, MC2ST filters whether to sample or not in order to avoid bias, i.e., when the first part of the sliding window has a relatively low anomaly score, then this step is skipped. Reconstruction Error Based Approach [69] Abnormal behavior detection is carried out by training the autoencoder. This is done by first partitioning the data into two subsets, normal and abnormal, and processing the abnormal subset by clustering. Then the normal subset is passed to the autoencoder for representation learning and iteration, during the iteration process, the samples are evaluated as normal or abnormal, and the normal samples obtained from the evaluation are re-input into the autoencoder for learning. Finally when the sample affiliation does not change, the abnormal score is calculated by using the reconstruction error of the self encoder as a scoring function. Temporal Masked Auto-Encoding (TMAE) [70] is an end-to-end fully unsupervised anomaly detection method.TMAE is inspired by Masked Auto-Encoding (MAE) [74] (MAE) inspired by the Masked Auto-Encoding (MAE). First the video foreground is recognized and Spatial-Temporal Cubes (STC) are constructed from consecutive image chunks, where STC represents video events. Half of the image blocks in the STC are then used to generate a mask along the temporal dimension, and the other half is used to train the VIT [75], thus predicting the image blocks with masks. Finally the sparsity of the occurrence of anomalous events is utilized to obtain the anomaly score. Self-trained Deep Ordinal Regression (SDOR) [71] is an end-to-end trainable video anomaly detection method.SDOR first anomaly scores a set of frames and generates pseudo-labels by two anomaly detectors in the initial module. Then, the set of frames is fed into an end-to-end anomaly score learner, and the anomaly score is updated by the result, and the pseudo-label of the set of frames is updated on the basis of the new anomaly score. Finally the anomaly detector of the initial module is trained for better detection results by iterating the above process. Generative Cooperative Learning (GCL) [72] GCL is a method to establish cross-supervision by generator and discriminator.GCL first uses deep feature extractor to convert video data into compact segment features and iteratively selects random samples among the features to train the GCL model.Then GCL generates pseudo-labels of the features through the generator so as to train the discriminator.Then the discriminator updates the pseudo-labels of the features to inversely train the generator to improve the accuracy of pseudo-labels through mutual iteration.Finally, the prediction result of the discriminator is used as the anomaly score. labeling accuracy, and finally the prediction result of the discriminator is used as the anomaly score. A Causal Inference Look (CIL) [73] is a method to address the confusion bias in the pseudo-label generation process.CIL first extracts features from a pre-trained CNN and generates initial pseudo-labels using the Random Forest algorithm. Then the initial pseudo-labels are used to train the CNN so as to re-generate the pseudo-labels, followed by an iteration of this step. Finally, the confounding effect of the pseudo-label generation process is addressed by intervening causal graphs, which makes the pseudo-label generation process not spuriously correlated with the iterative self-training process of the CNN by blocking the causal links. Coarse-to-Fine Pseudo-Labeling (C2FPL) [14] is a two-stage pseudo-label generation framework.C2FPL first generates a video-level pseudo-label for each video in the training set using a hierarchical clustering method in the coarse-pseudo-label generation stage. Then, in the fine-pseudo-label generation phase, segment-level pseudo-labels are generated for all segments in the training set by a statistical hypothesis testing method. Finally, in the anomaly detection phase, segment-level pseudo-labels are used to train an anomaly detector by supervised approach through which an anomaly score is obtained. Summary of Anomaly Detection Method Strategies can be seen in Table 5.

Table 5 Summary of Anomaly Detection Method Strategies

model classification		mould	timing	modeling strategy
supervised	on the basis of gap	Sabokrou et al. [33]	2018	Video Anomalous Behavior Detection Using Adversarial Training
		NNC [34]	2019	A two-stage anomaly detection framework
		Ionescu et al. [35]	2017	Converting the Video Abnormal Behavior Detection Problem to a Binary Classification Problem
		Ramachandra et al. [36]	2020	Learning Features and Distance Functions via Twin Networks
	on the basis of probability (math.)	SSMTL [44]	2023	Self-supervised multi-task joint learning
		STG-NF [45]	2023	time-space diagram normalized flow
		MocoDAD [46]	2023	denoising diffusion probability model
		PMAE [47]	2023	Semi-supervised Abnormal Behavior Detection Network Based on Probabilistic Memory Models
	on the basis of redevelopment	HF2-VAD [63]	2021	Integrated stream reconstruction and frame prediction for video anomaly detection
		SSPCAB [64]	2022	Learning to detect samples by convolutional layers with dilation filters
		STP [65]	2022	A predictive anomaly detection task based on spatio-temporal normality

		EVAL[61]	2023	Single-scene video anomaly localization framework
wholly unsupervised	sorter	Tudor et al. [66]	2017	Applying Text Authorship Authentication Techniques to Anomaly Detection
		MC2ST [68]	2018	A past-frame sampling method is proposed
		Li et al. [69]	2021	Divide the subset to train the autoencoder
		TMAE [70]	2022	Constructing Space-Time Cubes Training VIT
	false label	SODR [71]	2020	End-to-end trainable models
		GCL [72]	2022	Establishing cross-supervision through generators and discriminators
		CIL [73]	2022	Intervening causal diagrams to address confounding bias
		C2FPL [14]	2024	Two-stage pseudo-label generation framework

4 DATA SETS

Benchmark datasets play a crucial role in the study of video anomalous behavior detection problems [76], these datasets help to help detection models better understand the magnitude and scope of the occurrence of anomalous behaviors and provide a benchmarking platform to compare the performance of the models. For video anomalous behavior detection, this paper summarizes the currently available publicly available datasets in terms of their content, size, annotation style, source of video sequences, frame rate, and their characteristics, and Table 6 summarizes the datasets mentioned in this section.

Table 6 Summary of Abnormal Behavior Datasets

year	name (of a thing)	Number of videos			exceptios kind	Number of frames/video		Number of scenes	resolution (of a photo)
		Total	Training	Testing		Normal	Abnormal		
2008	Subway [77]	2	-	-	8	192548	16603	2	512 x 384
2013	UCSD ped1 [78]	70	34	36	5	9995	4005	1	158 x 238
2013	UCSD ped2 [78]	28	16	12	5	2924	1636	1	240 x 260
2013	CUHK Avenue [79]	37	16	21	6	26832	3820	1	640×360
2019	ShanghaiTech [80]	437	330	107	11	330 (Number of videos)	107 (Number of videos)	13	846×480
2020	Street Scene [23]	81	46	35	17	159341	43916	1	1280×720
2021	LAD2000 [81]	2000	1440	560	14	1300 (Number of videos)	700 (Number of videos)	1895	226 x 400
2022	UBnormal [82]	543	268	211	22	147887	89015	29	1280×720

4.1 LAD2000

LAD2000 dataset [81] is a large-scale anomalous behavior dataset, which consists of 2000 video sequences grouped into 14 anomalous event categories, including collision, crowd, destruction, fall, crash, fight, fire, water fall, injury, prowl, panic, theft, stampede, and violence, and each category consists of more than 100 video sequences, which are annotated by the corresponding video-level labels (anomalous/normal video, anomalous type) and frame-level labels (anomaly /normal video frames) comprising the corresponding annotated data.The LAD2000 is sourced from major public websites (YouTube, YouKuy, and Tencent Video), existing activity recognition databases (FCVID, holwood2, and YouTubeAction), and plazas recorded by the authors using a digital camera, as well as a number of normal activities and sudden occurrences of abnormal behaviors in the school. The frame rate was 25fps. LAD2000 includes a large number of visual scenes and real events suitable for testing anomalous behavior detection models and classification

models.LAD2000 provides us with a behaviorally distinct and categorically clear representation of events by discarding video sequences that are low-resolution and low-quality as well as those where the anomalous event is not completely unambiguous. In addition, LAD2000 records the entire process of the abnormal event from beginning to end and uses video sequences to represent the complete event, which is more advantageous for the detection of abnormal types such as fire and wandering, because these two types of events have a longer abnormal time, and they are labeled as anomalous frames from the beginning of the event, which can help the detection model to recognize the abnormal behavior and determine the category of the anomalies with a greater probability.

4.2 UCSD Pedestrian

UCSD Pedestrian dataset [78] is one of the most widely used datasets in video abnormal behavior detection models, which consists of UCSD Ped1 and UCSD Ped2, where Ped1 consists of 70 video sequences and Ped2 consists of 28 video sequences, and the two sub-datasets cover abnormal behaviors such as bicycling, skating, stroller, wheelchair, walking, and other, and consist of timestamps and pixel-level labels with corresponding Note that UCSD is derived from the motion of non-pedestrian objects as well as pedestrians on the sidewalk recorded by two stationary cameras fixed at higher positions in the school with a frame rate of 10fps. The UCSD mainly depicts some normal and abnormal pedestrian passages and the presence of transportation on the sidewalks, with a change in the video from sparse to crowded foot traffic due to the camera looking down.The difference between the Ped1 and Ped2 subdatasets is that Ped1 mainly includes clips of people walking towards and away from the camera, with some perspective distortion as well, whereas Ped2 mainly includes scenes of pedestrians moving while parallel to the camera.

4.3 CUHK Avenue

CUHK Avenue dataset [79] is the first dataset that introduces bags as a target for video anomalous behavior detection.The dataset consists of 47 video sequences covering anomalous behaviors such as paper throwing, bag throwing, children running and jumping, misdirection, and bags on grass.The annotated data consists of timestamps and pixel-level labels (in the form of bounding boxes).CUHK Avenue is derived from the campus roadway where people are recorded from a single camera in the school. scenes entering and exiting buildings at a frame rate of 25fps. The CUHK Avenue dataset was expanded from 15 video sequences in the early days to 47 video sequences.[83], and the authors proposed a sparse combination learning framework, which improved the detection speed from 140-150 frames/sec to 1000-1200 frames/sec. The CUHK Avenue dataset provides detailed pedestrian annotation information for each video sequence, including the pedestrian's location, motion trajectory, and behavioral actions. This annotation information can help researchers conduct effective model training and testing.

4.4 Street Scene

Street Scene [23] Street Scene is a large street scene dataset that contains a wide range of anomalous behaviors, consisting of 81 high-resolution video sequences of both anomalous behaviors such as jaywalking and illegal U-turns, as well as anomalous behaviors that did not occur in the simple training set such as walking pets and ticketing by staff, with annotated data consisting of bounding boxes delineating the anomalous areas and the numbering of the bounding boxes.Street Scene A two-lane street scene of a bike lane and sidewalk recorded from a camera fixed at an elevated location with a frame rate of 15 fps. The Street Scene dataset covers a variety of traffic behaviors as people move through the street, such as cars driving, turning, stopping, and pedestrians standing, walking, jogging, and pushing strollers, as well as bicyclists in bike lanes.Street Scene is shot overlooking the street at different times of the day during the summer months, so the dataset will have shading from trees and large vehicles, and changing shadows and moving backgrounds (e.g., flags and swinging trees in the wind), as influenced by the light and wind direction. There are also shifting shadows and moving backgrounds (e.g., flags flying in the wind and swinging trees) due to lighting and wind direction. The authors divided the training set to include only normal behaviors such as normal driving of vehicles and normal passing of people, and included 17 different types of abnormal behaviors in the test set.

4.5 Subway

Subway [77] is a dataset that captures the activities of people at subway entrances and exits. The dataset consists of two long video sequences of two different indoor scenes forming two sub-datasets. Anomalous behaviors include jumping, squeezing through a turnstile, cleaning a wall and walking in the wrong direction, among other anomaly types. The dataset has no spatially annotated data, only 85 anomalous events labeled by time.Subway is derived from two cameras at subway entrances and subway exits with a frame rate of 25 fps.Subway as an old dataset and suffers from low resolution and lack of spatial annotation, which is not conducive to model training.

4.6 ShanghaiTech

ShanghaiTech [80] is a multi-scene dataset. The dataset consists of 437 video sequences containing 13 different scenes. Abnormal behaviors include types of jostling, chasing, skating, cycling, and pushing carts on the sidewalks.ShanghaiTech is sourced from university sidewalk cameras. ShanghaiTech's 13 different anomalous scenes

were captured by multiple cameras with different viewpoints under different lighting conditions. The dataset can be viewed as 13 independent single-scene datasets, but it can also affect the training of the model due to the small number of anomalous events in a single scene.

4.7 UBnormal

UBnormal [82] is the first dataset with virtual scenes. The dataset consists of 543 video sequences containing 22 abnormal event types such as running, falling, fighting, sleeping, and crawling, which ensures that the training, test, and validation sets contain different types of abnormalities. Annotation data includes both frame-level and pixel-level. ubnormal comes from placing the generated virtual animated characters and objects in a real-world background with a frame rate of 30fps. UBnormal, as the first dataset to provide a validation set for model tuning, ensures the possibility of model tuning without overfitting the test set during testing of the model. This point avoids the two possibilities of traditional debugging models (1) tuning a model based on test data and inherently overfitting it (2) tuning by hyperparameters which may lead to sub-optimal results.

5 EVALUATION CRITERIA

Abnormal behavior is generally defined differently depending on the scenario, fighting behavior is an abnormal event that becomes normal in the scenario of a boxing match scene. Although the annotated data of video sequences are only abnormal and normal in nature, ambiguities such as these can exist. A good evaluation criterion for abnormal behavior should be able to describe as much as possible the qualitative performance of an abnormal behavior detection model in practice, based on the specifics of different scenarios. Calculation of the evaluation metrics first requires the selection of a threshold value. Samples with abnormal scores below the threshold are considered normal and vice versa. This can be represented using a confusion matrix, where TP, FN, FP, and TN denote the number of correctly detected abnormal samples, the number of abnormal samples misdetected as normal, the number of normal samples misdetected as abnormal, and the number of normal samples correctly detected, respectively.

5.1 Frame-Level and Pixel-Level Standards

Researchers have made extensive use of frame-level and pixel-level criteria [84] to evaluate the performance of detection models [34,46,58]. The frame-level criterion is to count all detected frames with abnormal pixels as positive frames and count the remaining other frames as normal frames, and then determine which detected frames are true positives and which are false positives by Ground-Truth to compute the True Positive Rate (TPR) and False Positive Rate (FPR) at a given abnormality scoring threshold. False Positive Rate, FPR). The frame-level criterion does not use spatial localization, which means that detected anomalous pixels are considered correctly detected even if they are different from Ground-Truth.

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{2}$$

$$\mathrm{FPR}=\frac{\mathrm{FP}}{\mathrm{FP+TN}} \tag{3}$$

The frame level criterion is based on the fact that given the tth frame of the test video corresponding to the pixel anomaly scoring map *St*, the frame is detected as anomalous if $\sum p\ [St\ (p)\geq\Gamma] \geq 1$, where *P* is the total number of pixels and Γ is the anomaly scoring threshold, otherwise the frame is detected as normal. If $\sum p\ [At\ (p)==1] \geq 1$, the frame is counted as a true positive frame, where *At* stands for Ground-Truth. if $\sum p\ [At\ (p)==1] == 0$, the frame is counted as a false positive frame.

$$\mathrm{TP+FN}=\sum_{t=1}^{T}\left(\sum_{p}[A^{t}(p)=1]\geq 1\right) \tag{4}$$

$$\mathrm{FP+TN}=\sum_{t=1}^{T}\left(\sum_{p}[A^{t}(p)=1]=0\right) \tag{5}$$

The pixel-level criterion incorporates the spatial localization of anomalies by counting true-positive and false-positive frames instead of true-positive and false-positive anomaly regions. A detected anomalous pixel is considered a true positive if it accounts for at least 40% of the Ground-Truth. Other detected anomalous pixels that do not overlap with Ground-Truth are ignored.

$$\begin{aligned}\mathrm{TP+FN}=\sum_{t=1}^{T}\{&\sum_{P}[S^{t}(p)\geq\Gamma\cap A^{t}(p)]\geq 0.4\\ &\cap\sum_{p}[A^{t}(p)=1]\cap\sum_{p}p[A^{t}(p)=1]\geq 1\}\end{aligned} \tag{6}$$

The pixel level criterion is based on the fact that given the tth frame of the test video corresponding to the pixel anomaly scoring map *S*t, the frame is counted as a true positive frame if $\sum p[St\ (p)\geq\Gamma\cap At\ (p)] \geq 0.4\cap\sum p[At\ (p)==1]$ and $\sum p[At\ (p)==1]=1$. If $\sum p[St\ (p)\geq\Gamma] \geq 0$ and $\sum p[At\ (p)==1]==0$, the frame is counted as a false positive frame. False-positive frames are computed without any frames containing the same pixel values as Ground-Truth [84]. This addition leads to two results, the first is that if the predicted frame contains only one anomalous pixel with the same value as Ground-Truth, the frame cannot be counted as a false positive frame, and the second is that if more than one

region in the frame is predicted to be anomalous, but as long as the frame does not contain any anomalous pixel with the same value as Ground-Truth, the frame is counted as a false positive frame. Based on this addition a simple post-processing of the anomaly scoring map can be performed [23] that makes the pixel-level criterion equivalent to the frame-level criterion by ignoring the tightness or looseness of the localization. This is done by marking all the pixels of the frame as anomalous for any frame in which anomalous pixels are detected, a move that ensures that the frame passes Ground-Truth's 40% threshold for detection, whereas frames in which no anomalies are detected do not further increase the FPR, and therefore do not have an impact on the detection results, and the use of a pixel-level criterion is recommended since frame-level criteria do not allow for evaluating the implementation of spatial localization [2].

5.2 Dual Pixel Standard

To overcome the limitations of the frame-level and pixel-level criteria, the two-pixel criterion [85] adds a new constraint P1 to the original pixel-level standard, i.e., the pixels detected as anomalous must first satisfy that the pixels detected as anomalous account for at least 10% of the Ground-Truth in order to perform the post-processing operation of marking all the images of the frame as anomalous, and then detecting it by the Ground-Truth's threshold of 40%, which results in the counting of the frame as a true positive frame.

$$\mathrm{P_1}=\sum_{P}[S^t(p)\geq\Gamma\bigcap A^t(p)]\geq 0.1 \tag{7}$$

While the two-pixel criterion was able to address to some extent the problem of post-processing methods over-increasing the number of true-positive frames, it also resulted in other outcomes. For example, (1) the frame is detected to have multiple true anomalous pixels, but does not reach 10% of Ground-Truth and cannot be counted as a true positive frame. (2) The frame was detected to have multiple false anomaly pixels and reached 10% of Ground-Truth and was counted as a true positive frame. (3) The frame is detected to have both true anomalous pixels and false anomalous pixels that add up to 10% of Ground-Truth and is counted as a true positive frame. All three results lead to the inability to correctly count the true-positive and false-positive frames. To solve this problem, Lu et al. [83] proposed Intersection Over Union (IOU) applied to the CUHK Avenue dataset to constrain the tightness and looseness of spatial localization. However, this method is not effective in solving the problem of accurate counting of positive and false-positive frames, and IOU cannot be applied to other datasets due to the differences in annotation formats of different datasets.

5.3 Regionally based Standards

Region-based criteria [86] By testing the performance of a model in a more realistic way, the evaluation criteria should take into account any expected ambiguities and biases that would occur in the dataset. In order to solve the problems arising from traditional criteria, region-based criteria take two steps, the first is to achieve spatial localization by proposing a loose IOU mechanism to account for the ambiguity between the anomaly detection results and Ground-Truth, and the second is to count the detected regions as either true-positive or false-positive regions, such that any frame can be more than just a single counted as either a true positive frame or false positive frame. The Region-Based Detection Rate (RBDR) and the FPR of each frame are calculated from the number of true-positive regions and the number of false-positive regions, and the Area under the ROC curve (AUC) is further calculated.

$$\mathrm{RBDR}=\frac{\mathrm{NTP}}{\mathrm{TAR}} \tag{8}$$

The number of True Positive (NTP) is calculated by the IOU mechanism, where *Dt* is the detected anomalous region in frame t, G_t^i is the ith Ground-Truth labeled region in frame t, *Nt* is the total number of Ground-Truth labeled regions in frame t, and β is the threshold value. The region-based criterion considers that an accurate NTP should represent all detected truly anomalous regions, and the detected anomalous regions are IOU'd with the Ground-Truth labeled regions, and at least one of the ratios is greater than β, i.e., the detected regions are counted as NTPs.

5.4 Trajectory-based Criteria

Track-Based Criteria [86] As with the region-based criterion, the same two steps are used to count true-positive or false-positive tracks and false-positive regions, and to calculate the Track-Based Detection Rate (TBDR) and the FPR, which in turn calculates the AUC.

$$\mathrm{TBDR}=\frac{\mathrm{NTPT}}{\mathrm{NAT}} \tag{9}$$

The calculation of True Positive Tracks (NTPT) requires the participation of Ground-Truth labeled anomalous tracks L_k, *Lk* refers to a set of Ground-Truth labeled anomalous regions in a sequence of consecutive frames, while L_k is derived by assuming, under general conditions, that *the k* in G_t^k labels a particular Ground -Truth labeled anomalous regions in frame t. The set L_k formed by concatenating the *k* regions in a sequence of consecutive frames is the Ground-Truth labeled anomalous trajectories. *nk* is the total number of detected anomalous trajectories, and α is the threshold value. The anomalous region *Dt* in the detected anomalous trajectories is IOU with the anomalous region G_t^k in *Lk*, and at least one ratio is greater than α, then the detected trajectories are counted as NTPT.

The FPR is calculated in the same way and is derived from the ratio of NFP to the total number of frames, which is used to evaluate the model performance. In addition, since both new criteria involve calculating the number of false positive regions, and there is more than one false positive region in each frame, the maximum FPR for both criteria may exceed 1.0, and the ROC curve is evaluated by plotting the ROC curve of FPR and further calculating the AUC [86], with a range of [0, 1.0]. It is feasible to use AUC to evaluate FPR [1], because the quality of false positives counted by different evaluation criteria cannot be captured without visual inspection, AUC is needed to provide qualitative analysis and visualization. In addition the use of two new criteria, distance-based and trajectory-based, requires datasets with Ground-Truth labeled anomalous areas and anomalous trajectories, and the authors of the area-based criterion provided annotations for both the UCSD Pedestrian, CUHK Avenue, and Street Scene datasets [86].

6 Comparative Analysis of Experiments with Different Algorithms

UCSD, CUHK Avenue, ShanghaiTech, Subway and XD-Violence are used as common datasets to compare the performance of different models more objectively, and the evaluation metrics data are quoted from the corresponding papers of each model. The evaluation metrics use frame-level, pixel-level, area-based and trajectory-based criteria for AUC and Average-Precision (AP) obtained from Precision Recall (PR) curves.

Table 7 Summary of Performance Comparison of Abnormal Behavior Detection Models

categorization		mould	Ped2	Avenue	ShanghaiTech	Subway
			Pixel/Frame	Frame	RBDC/TBDC/Frame	Extrance/Exit
supervised	on the basis of redevelopment	NNC [34]	97.80%/-	88.9%	-/-/-	-/-
		Ionescu et al. [35]	-/-	90.40%	-/-/-	-/-
		Ramachandra et al.[36]	94.00%/-	87.20%	-/-/-	-/-
	on the basis of probability (math.)	SSMTL [44]	-/-	92.50%	47.10%/85.60%/92.50%	-/-
		STG-NF [45]	-/-	-	52.10%/82.40%/95.90%	-/-
		MocoDAD [46]	-/-	89.00%	-/-/-	-/-
		PMAE [47]	95.90%/-	-	-/-/-	-/-
	on the basis of redevelopment	EVAL [61]	-/-	86.02%	59.21%/89.44%/76.63%	-/-
		HF2-VAD [63]	99.30%/-	90.30%	-/-/-	-/-
		SSPCAB [64]	-/-	92.90%	45.45%/84.50%/89.50%	-/-
		STP [65]	98.90%/-	90.10%	51.60%/84.60%/86.20%	-/-
wholly unsupervised	sorter	Tudor et al. [66]	-/82.20%	80.60%	-/-/-	70.60%/85.70%
		MC2ST [68]	-/85.80%	84.40%	-/-/-	71.70%/93.10%
		Li et al. [69]	-/-	-	-/-/-	-/-
		TMAE [70]	-/94.10%	89.80%	-/-/71.40%	-/-
	false label	SDOR [71]	-/83.20%	-	-/-/-	88.10%/92.70%
		GCL [72]	-/-	-	-/-/78.93%	-/-
		CIL [73]	-/98.70%	90.30%	-/-/-	91.30%/97.60%
		C2FPL [14]	-/-	-	-/-/-	-/-
legacy unsupervised	-	AST-AE [87]	-/96.70%	87.80%	-/-/-	-/-
		STC-Net [88]	-/98.10%	89.80%	-/-/73.80%	-/-
		STM-AE [89]	-/97.40%	86.70%	-/-/-	-/-
		Deng et al. [90]	-/97.60%	90.90%	-/-/78.80%	-/-
		Shi et al. [91]	-/96.60%	85.20%	-/-/-	-/-
		Le et al. [92]	-/98.90%	89.70%	-/-/75.00%	-/-

The following conclusions can be drawn by comparing the experimental results:

From the comparison results of the UCSD Ped2 dataset, as shown in Table 7, the FPN-based method [35] reached 97.80% of pixel-level AUC, PMAE [47] achieves 95.90% pixel-level AUC, and HF2-VAD [63] achieves 99.30% pixel-level AUC for HF-VAD. It can be seen that the performance of the video abnormal behavior detection algorithms on the UCSD dataset continues to improve, with HF2-VAD outperforming the other methods, and its 99.30% pixel-level AUC will result in the other algorithms not being able to demonstrate a significant increase in performance by the frame-level AUC criterion on the UCSD Ped2 dataset. Further comparisons of model performance can be made with other datasets.

From the comparison results of the CUHK Avenue dataset, as shown in Table 7, the FPN-based method [35] reached 88.90% of frame-level AUC, SSMTL [44] achieves 92.50% frame-level AUC, and SSPCAB[64] The frame-level AUC of SSPCAB reaches 92.90%.SSPCAB has the best performance in the CUHK Avenue dataset, which indicates that the reconstruction-based method helps to improve the anomaly detection by using the autoencoder and generative adversarial network to process the features. The similarity of these three methods is that all of them pay attention to the feature extraction process, which indicates that feature extraction and processing are in a key position in the process of video anomalous behavior detection.

From the comparison results of Shanghaitech dataset, as shown in Table 7, SSMTL [44] has a frame-level AUC of 92.50%, EVAL [58] RBDC reaches 89.44%,and TBDC reaches 59.21%.SSMTL performs best under the frame-level criterion, and EVAL performs best under the region-based and trajectory-based criteria. The RBDC scores of the five

methods are significantly lower than the TBDC, which is because the region-based criterion is more difficult to determine the anomalous region by detecting the anomalous region with Ground-Truth for IOUs and counting it as NTP when it is larger than the threshold, compared to the trajectory-based criterion which determines the anomalous region by the anomalous action trajectory.EVAL focuses on modeling through the correlation information of the action, and thus obtains the highest TBDC score.

Table 8 Comparison of Fully Unsupervised and Traditional Unsupervised

	Methods	Ped2	Avenue	ShanghaiTech
wholly unsupervised	Tudor et al. [66]	82.20%	80.60%	-
	MC2ST [68]	85.80%	84.40%	-
	TMAE [70]	94.10%	89.80%	71.40%
	SDOR [71]	83.20%	-	-
	GCL [72]	-	-	78.93%
	CIL [73]	98.70%	90.30%	-
legacy unsupervised	AST-AE [87]	96.60%	85.20%	68.80%
	STC-Net [88]	96.70%	87.80%	73.10%
	STM-AE [89]	98.10%	89.80%	73.80%
	Deng et al. [90]	98.90%	89.70%	75.00%
	Shi et al. [91]	97.60%	90.90%	78.80%
	Le et al. [92]	97.40%	86.70%	73.60%

In the fully unsupervised approach, as shown in Table 8, CIL [73] achieves 98.70% and 90.30% frame-level AUC on the UCSD Ped2 dataset and CUHK Avenue dataset, respectively, and GCL [72] achieves a frame-level AUC of 78.93% on the ShanghaiTech dataset. Among the traditional unsupervised methods, the bi-directional interpolation-based method [90] reached 98.90% frame-level AUC on the UCSD Ped2 dataset, the frame reconstruction-based method [91] on the CUHK Avenue dataset and ShanghaiTech dataset reached 90.90% and 78.80% of the frame-level AUC of the frame-level AUC, respectively. From the experimental results the fully unsupervised method is slightly lower than the unsupervised method on the UCSD Ped2 dataset and slightly higher than the unsupervised method on the CUHK Avenue dataset and the ShanghaiTech dataset, which illustrates that the fully unsupervised method is not inferior to the traditional unsupervised method in terms of performance.

In the two sub-datasets of Subway, as shown in Table 7, the classifier-based methods [68] reached 71.70% and 93.10% frame-level AUCs, respectively, and the pseudo-labeling-based method CIL [73] reaches 91.30% and 97.60% of frame-level AUC, which is the best performance in the Subway dataset. This indicates that CIL effectively generates pseudo-labels through Random Forest and CNN for the training of the detection model, and overcomes the confusion effect generated by other methods in generating pseudo-labels. The classifier-based approach does not show significant advantages over the pseudo-label-based approach, indicating that the direct training of classifiers by unlabeled features does not yield good results.

7 DIRECTIONS FOR FUTURE RESEARCH

7.1 Virtual Synthetic Data Sets

For the training and testing of video anomalous behavior detection algorithms, on the one hand, scene- and action-rich datasets are needed to complete the training, and on the other hand, it is more costly to manually collect and label the datasets.The proposal of UBnormal dataset provides a good idea to solve this problem through the generation of virtual animated characters and objects placed in real-world backgrounds, and provides both frame-level and pixel-level Ground-Truth. There is much more to consider for virtual synthetic datasets. Virtual synthetic datasets can focus more on diversity and complexity by simulating a variety of different abnormal behavior scenarios, including different actions, environments, lighting conditions, etc. This will help improve the generalization ability of the model so that it can detect various abnormal behaviors more accurately in the real world. The virtual synthetic dataset can also be further augmented with feature modalities by adding subtitles and sounds, which in turn solves the problem of sparse multimodal datasets.

7.2 Multimodal Large Models

Multimodal macromodel as a deep learning-based natural language processing model can be used to improve the performance of video anomalous behavior detection algorithms [93]. In the future, multimodal macromodels can be used to learn the video content to help understand the contextual information in the video, including scenes, dialogues and action sequences, and generate textual descriptions, which can help the detection algorithms to understand abnormal behaviors as well as timing information in the video. Other modal features such as appearance, motion and audio can also be further fused on top of the textual description to further improve the accuracy of the detection algorithm. In addition to this, the textual descriptions generated by the multimodal macromodel based on the video can also be applied in the automatic annotation of the dataset to improve the performance by the detection algorithms getting trained and tested more accurately.

7.3 Lightweight Models

The deployment in edge devices should be considered first in the future application of video abnormal behavior detection. On the one hand, most of the existing detection algorithms over-pursuing the performance while generating a large amount of operating costs and are unable to perform online real-time detection due to offline training, and on the other hand, edge devices have limited computational resources, and high robustness and real-time are the basic needs of detection algorithms for edge device applications.

The development of lightweight models can consider designing lightweight feature extraction modules to reduce computational complexity, or through pruning techniques[94] to remove redundant connections and parameters in the network to reduce the size and computation of the model while maintaining the performance of the model. Knowledge distillation[95] and model compression[96] to train a large model to guide lightweight model learning, thus improving the performance of the lightweight model.

8 CONCLUSION

This paper combs the tasks of abnormal behavior detection around video abnormal behavior detection methods, data sets and evaluation criteria. Starting from the detection methods, this paper first describes three supervised methods, namely distance based, probability based, and reconstruction based. At the same time, it introduces some new developments in completely unsupervised methods. Secondly, it systematically explains the data sets and evaluation criteria commonly used in video anomaly detection, and conducts experimental comparison and performance evaluation of the main methods from the data sets and evaluation criteria, and analyzes some possible factors that determine the performance of the model. Finally, this paper briefly describes the problems and trends that need to be concerned about in the future research direction of abnormal behavior detection methods.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported in part by the Social Science Fund of Guangxi under Grant 23FTQ005.

REFERENCES

[1] Yao Huiling, Xing HU. A survey of video violence detection. Cyber-Physical Systems. 2023, 9(1): 1-24.
[2] S.Roshan, G. Srivathsan, K. Deepak, S. chandrakala, et al. Violence detection in automated video surveillance: Recent trends and comparative studies. The Cognitive Approach in Cloud Computing and Internet of Things Technologies for Surveillance Tracking Systems. 2020: 157-171.
[3] Bharathkumar Ramachandra, Michael Jones, Ranga Raju Vatsavai. A survey of single-scene video anomaly detection. IEEE on pattern analysis and machine intelligence, 2020, 44(5): 2293-2312.
[4] Kamal Kant Verma, Brij Mohan Singh, Amit Dixit. A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. International Journal of Information Technology, 2019: 1-14.
[5] Hu Chunyu, Chen Yiqiang, Hu Lisha, et al. A novel random forests based class incremental learning method for activity recognition. Pattern Recognition, 2018, 78: 277-290.
[6] Xiao Qinkun, Song Ren. Action recognition based on hierarchical dynamic Bayesian network. Multimedia Tools and Applications, 2018, 77(6): 6955-6968.
[7] Sok Pichleap, Xiao Ting, Azeze Yohannes, et al. Activity Recognition for Incomplete Spinal Cord Injury Subjects using Hidden Markov Models. IEEE Sensors Journal, 2018, 18(15): 6369-6374.
[8] Bilal M'hamed Abidine, Lamya Fergani, Belkacem Fergani, et al. The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition. Pattern Analysis and Applications, 2018, 21(1): 119-138.
[9] Sun Xiaohu, Yu Axiang, Shen Xulin, et al. Abnormal Behavior Recognition Based on Hybrid Attention Mechanism. Computer Engineering and Applications, 2023, 59(5): 140-147.
[10] Liu Yang, Liu Jiang, Zhao Mengyang, et al. Learning appearance-motion normality for video anomaly detection. Proceedings of the 2022 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2022: 1-6.
[11] Li Daoheng, Nie Xiushan, Li Xiaofeng, et al. Context-related video anomaly detection via generative adversarial network. Pattern Recognition Letters, 2022, 156: 183-189.
[12] Lee Joo Yeon, NAM Woo Jeoung, Lee Seong Whan. Multi-contextual predictions with vision transformer for video anomaly detection. Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2022: 1012-1018.
[13] Tarik Alafif, Anas Hadi, Manal Allahyani, et al. Hybrid classifiers for spatio-temporal abnormal behavior detection, tracking, and recognition in massive Hajj crowds. Elecronics, 2023, 12(5): 1165.

[14] Anas Al-lahham, Nurbek Tastan, Zaigham Zaheer, et al.A Coarse-to-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection. arXiv preprint arXiv: 2310.17650, 2024.
[15] WONG Sebastien C, Stamatescu Victor, GATT Adam, et al. Track Everything: Limiting Prior Knowledge in Online MultiObject Recognition. IEEE Transactions on Image Processing, 2017, 26(10): 4669-4683.
[16] Kong Xiangjie, Ma Kai, Hou Shen, et al. Human interactive behavior: A bibliographic review. IEEE Access, 2018, 7: 4611-4628.
[17] Bahram Mohammadi, Mahmood Fathy, Mohammad Sabokrou. Image/video deep anomaly detection: A survey. arXiv preprint arXiv: 2103.01739, 2021.
[18] Xu Dan, Yan Yan, Ricci Elisa, et al. Detecting anomalous events in videos by learning deep representations of appearance and motion. Computer Vision and Image Understanding, 2017, 156: 117-127.
[19] Sabokrou Mohammad, Fathy Mahmood, Hoseini Mojtaba, et al. Realtime anomaly detection and localization in crowded scenes. Proceedings of the 2015 IEEE conference on computer vision and pattern recognition workshops. Piscataway: IEEE, 2015: 56-62.
[20] Sabokrou Mohammad, Fayyaz Mohsen, Fathy Mahmood, et al. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding, 2018, 172: 88-97.
[21] Kim Junbong, Jeong Kwanghee, Choi Hyomin, et al.GAN-based anomaly detection in imbalance problems. Proceedings of the 2020 European Conference on Computer Vision (ECCV). Berlin: Springer, 2020: 128-145.
[22] Tang Yao, Zhao Lin, Zhang Shanshan, et al. Integrating prediction and reconstruction for anomaly detection. Pattern Recognition Letters, 2020, 129: 123-130.
[23] Ramachandra Bharathkumar, Jones Michael J.Street scene: A new dataset and evaluation protocol for video anomaly detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2569-2578.
[24] Alhothali Areej, Balabid Amal, Alharthi Reem, et al. Anomalous event detection and localization in dense crowd scenes. Multimedia Tools and Applications, 2023, 82(10): 15673-15694.
[25] Liu Yang, Yang Dingkang, Fang Gaoyun, et al. Stochastic video normality network for abnormal event detection in surveillance videos. Knowledge-Based Systems, 2023, 280: 110986.
[26] Alhothali Areej, Balabid Amal, Alharthi Reem, et al. Anomalous event detection and localization in dense crowd scenes. Multimedia Tools and Applications, 2023, 82(10): 15673-15694.
[27] Sattarzadeh Sam, Sudhakar Mahesh, Plataniotis Konstantinos N. SVEA: asmall-scale benchmark for validating the usability of posthoc explainable AI solutions in image and signal recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 4158-4167.
[28] Zhan Cheng, Hu Han, Wang Zhi, et al. Unmanned aircraft system aided adaptive video streaming: A joint optimization approach. IEEE Transactions on Multimedia, 2019, 22(3): 795-807.
[29] Zhang Zhe, MA Shiyao, Yang Zhaohui, et al. Robust Semisupervised Federated Learning for Images Automatic Recognition in Internet of Drones. IEEE Internet of Things Journal, 2022, 10(7): 5733-5746.
[30] Jessie James P. Suarez, Prospero C. Naval Jr, A survey on deep learning techniques for video anomaly detection. arXiv preprint arXiv: 2009.14146, 2020.
[31] Cai Yiheng, Liu Jiaqi, Guo Yayun, et al. Video anomaly detection with multi-scale feature and temporal information fusion. Neurocomputing, 2021, 423: 264-273.
[32] Xinyang Feng, Dongjin Song, Yuncong Chen, et al. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 5546-5554.
[33] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, et al. Adversarially learned one-class classifier for novelty detection. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018: 3379-3388.
[34] Ionescu Radu Tudor, Smeureanu Sorina, Popescu Marius, et al. Detecting abnormal events in video using narrowed normality clusters. Proceedings of the 2019 IEEE winter conference on applications of computer vision (WACV). Piscataway: IEEE, 2019: 1951-1960.
[35] Ionescu Radu Tudor, Khan Fahad Shahbaz, Georgescu Mariana-Iuliana, et al. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 7842-7851.
[36] Ramachandra Bharathkumar , Jones Michael J, Vatsavai R. Learning a distance function with a Siamese network to localize anomalies in videos. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2598-2607.
[37] Gong Suming, Chen Ying. Video Action Recognition Based on Spatio-Temporal Feature Pyramid Module. Journal of Frontiers of Computer Science and Technology, 2022, 16(9): 2061-2067.
[38] Erkan Şengönül, Refik Samet, Qasem Abu Al-Haija, et al. An analysis of artificial intelligence techniques in surveillance video anomaly detection: A comprehensive survey. Applied Sciences, 2023, 13(8): 4956.
[39] García González Jorge, Molina-Cabello Miguel A, Luque-Baena Rafael M, et al. Deep autoencoder architectures for foreground object detection in video sequences based on probabilistic mixture models. Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2020: 3199-3203.

[40] Manoj Kumar Sharma, Debdoot Sheet, Prabir Biswas. Spatiotemporal deep networks for detecting abnormality in videos. Multimedia Tools and Applications, 2020, 79(15): 11237-11268.
[41] P. Kuppusamy, V.C. Bharathi. Human abnormal behavior detection using CNNs in crowded and uncrowded surveillance A survey. Measurement: Sensors, 2022, 24: 100510.
[42] Jaechul Kim, Kristen Grauman, Observe locally, infer globally: a space-time MRF for detecting abnormal activities with in cremental updates. Proceedings of the 2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2009: 2921-2928.
[43] Wu Shandong, Brian E. Moore, Mubarak Shah, Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. Proceedings of the 2010 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2010: 2054-2060.
[44] Antonio Barbalau, Radu Tudor Ionescu, Mariana-Iuliana Georgescu, et al. SSMTL++: Revisiting self-supervised multi-task learning for video anomaly detection. Computer Vision and Image Understanding, 2023, 229: 103656.
[45] Or Hirschorn, Shai Avidan. Normalizing flows for human pose anomaly detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 1354 5-13554.
[46] Alessandro Flaborea, Luca Collorone, Guido Maria, et al. Multimodal Motion Conditioned Diffusion Model for Skeleton-based Video Anomaly Detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 10318-10329.
[47] Xiao Jinsheng, Guo Haowen, Xie Hongang, et al. Probabilistic Memory Auto-encoding Network for Abnormal Behavior Detection in Surveillance Videos. Journal of Software, 2023, 34(9): 4362-4377.
[48] Alexander Gepperth, A new perspective on probabilistic image modeling. Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2022: 1-10.
[49] Jonathan Ho, Ajay Jain, Pieter Abbeel, Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020, 33: 6840-6851.
[50] Xiao Jinsheng, Shen Mengyao, Jiang Mingjun, et al. Monitoring vide oabnormal behavior detection with fusion package attentionmechanism. Acta Automatica Sinica, 2022, 48(12): 2951-2959.
[51] Liu Yang, Liu Jing, Lin Jieyu, et al. Appearance-motion united auto encoder framework for video anomaly detection. IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 69(5): 2498-2502.
[52] Liu Wenrui, Chang Hong, Ma Bingpeng, et al. Diversity-measurable anomaly detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 12147-12156.
[53] Huang Xiangyu, Zhao Caidan, Wu Zhiqiang. Updated version: A Video Anomaly Detection Framework based on Appearance-Motion Semantics Representation Consistency. arXiv preprint arXiv: 2303.05109, 2023.
[54] Andrea Borghesi, Andrea Bartolini, Michele Lombardi. Anomaly detection using a convolutional winner-take-all auto encoder. Proceedings of the 2017 British machine vision conference. Berlin: Springer, 2017: 1-12.
[55] Gong Dong, Liu Lingqiao, Le Vuong, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 1705-1714.
[56] Sun jingbo, Ji jie. Memory-augmented deep autoencoder model for pedestrian abnormal behavior detection in video surveillance. Infrared and Laser Engineering, 2022, 51(6): 20210680.
[57] Liu Chengming, Xue Ran, Shi Lei, et al. The gating self-attention mechanism and GAN integrated video anomaly detection. Journal of Image and Graphics, 2022, 27(11): 3210-3221.
[58] Chen Dongyue, Yue Lingyi, Chang Xingya, et al. NM-GAN: Noise-modulated generative adversarial network for video anomaly detection. Pattern Recognition, 2021, 116: 107969.
[59] Zhang Qianqian, Feng Guorui, Wu Hanzhou. Surveillance video anomaly detection via non-local U-Net frame prediction. Multimedia Tools and Applications, 2022, 81(19): 27073-27088.
[60] Huang Chao, Wu Zhihao, Wen Jie, et al. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. IEEE Transactions on Industrial Informatics, 2021, 18(8): 5171-5179.
[61] Ashish Singh, Michael J. Jones, Erik G. Learned-Millerv. Eval: Explainable video anomaly localization. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 18717-18726.
[62] Zhang Hongmin, Zhuang Xu, Zheng Jingtian, et al. Optimizing Human Abnormal Behavior Detection Method of YOLO Network. Computer Engineering and Applications, 2023, 59(7): 242-249.
[63] Liu Zhian, Nie Yonngwei, Long chengjiang, et al. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 13588-13597.
[64] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, et al. Self-supervised predictive convolutional attentive block for anomaly detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 13576-13586.
[65] Yassine Naji, Aleksandr Setkov, Angélique Loesch, et al. Spatio-temporal predictive tasks for abnormal event detection in videos. Proceedings of the 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Piscataway: IEEE, 2022: 1-8.

[66] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, et al. Unmasking the abnormal events in video. Proceedings of the 2017 IEEE international conference on computer vision. Piscataway: IEEE, 2017: 2895-2903.
[67] Wu Kaijun, Huang Tao, Wan Dicong, Bai Chenshuai, et al. Research Progress of Video Anomaly Detection Technology. Journal of Frontiers of Computer Science and Technology, 2022, 16(3): 529-540.
[68] Liu Yusha, Li Chunliang, Barnabás Póczos. Classifier Two Sample Test for Video Anomaly Detections. Proceedings of the 2018 British machine vision conference. Berlin: Springer, 2018: 71.
[69] Li Tangqing, Wang Zheng, Liu Siying, et al. Deep unsupervised anomaly detection. Proceedings of the IEEE/CVF winter conference on applications of computer vision. Piscataway: IEEE, 2021: 3636-3645.
[70] Hu Jingtao, YU Guang, Wang Siqi, et al. Detecting Anomalous Events from Unlabeled Videos via Temporal Masked Auto-Encoding. Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2022: 1-6.
[71] Pang Guansong, Yan Cheng, Shen Chunhua, et al. Self-trained deep ordinal regression for end-to-end video anomaly detection. Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition. Piscataway: 2020: 12173-12182.
[72] M. Zaigham Zaheer, Arif Mahmood, M. Haris Khan, et al. Generative cooperative learning for unsupervised video anomaly detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 14744-14754.
[73] Lin Xiangru, Chen Yuyang, Li Guanbin, et al. A causal inference look at unsupervised video anomaly detection. Proceedings of the 2022 AAAI Conference on Artificial Intelligence. Menlo Park: AAAI, 2022, 36(2): 1620-1629.
[74] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked Autoencoders Are Scalable Vision Learners. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 16000-16009.
[75] Yuan Hongchun, Cai Zhenyu, Zhou Hui, et al. Transanomaly: Video anomaly detection using video vision transformer. IEEE Access, 2021, 9: 123977-123986.
[76] Kong Xiangjie, Liu Xiaoteng, Behrouz Jedari, et al. Mobile crowd sourcingin smart cities: Technologies, applications, and future challenges. IEEE Internet of Things Journal, 2019,6(5): 8095-8113.
[77] Amit Adam, Ehud Rivlin, Ilan Shimshoni, et al. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(3): 555-560.
[78] Li Weixin, Vijay Mahadevan, Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(1): 18-32.
[79] Lu Cewu, Shi Jianping, Jia Jiaya. Abnormal event detection at 150 fps in matlab. Proceedings of the 2013 IEEE international conference on computer vision. Piscataway: IEEE, 2013: 2720-2727.
[80] Liu Wen, Luo Weixin, Lian Dongze, et al. Future frame prediction for anomaly detection: a new baseline. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2018: 6536-6545.
[81] Wan Boyang, Jiang Wenhui, Fang Yuming, et al. Anomaly detection in video sequences: A benchmark and computational model. IET Image Processing, 2021, 15(14): 3454-3465.
[82] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, et al. Ubnormal: New benchmark for supervised open-set video anomaly detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 20143-20153.
[83] Lu Cewu, Shi Jianping, Wang Weiming, et al. Fast abnormal event detection. International Journal of Computer Vision, 2019, 127: 993-1011.
[84] Wang Shu, Miao Zhenjiang. Anomaly detection in crowd scene. Proceedings of the 10th IEEE International Conference on Signal Processing Proceedings. Piscataway: IEEE, 2010: 1220-1223.
[85] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, et al. Deep cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. IEEE Transactions on Image Processing, 2017, 26(4): 1992-2004.
[86] Bharathkumar Ramachandra, Michael J. Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. Proceedings of the 2020 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2020: 2569-2578.
[87] Liu Yang Li Shuang, Liu Jing, et al. Learning attention augmented spatial-temporal normality for video anomaly detection. Proceedings of the 3rd International Symposium on Smart and Healthy Cities (ISHC). Piscataway: IEEE, 2021: 137-144.
[88] Zhao Mengyang, Liu Yang, Liu Jing, et al. Exploiting spatial-temporal correlations for video anomaly detection. Proceedings of the 26th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2022: 1727-1733.
[89] Liu Yang, Liu Jing, Zhao Mengyang, et al. Learning appearance-motion normality for video anomaly detection. Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2022: 1-6.
[90] Deng Hanqiu, Zhang Zhaoxiang, Zou Shihao, et al. Bi-Directional Frame Interpolation for Unsupervised Video Anomaly Detection. Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (CVPR). Piscataway: IEEE, 2023: 2634-2643.

[91] Shi Chenrui, Sun Che, Wu Yuwei, et al. Video anomaly detection via sequentially learning multiple pretext tasks. Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (CVPR). Piscataway: IEEE, 2023: 10330-10340.
[92] Viet-Tuan Le, Yong-Guk Kim. Attention-based residual autoencoder for video anomaly detection. Applied Intelligence, 2023,53(3): 3240-3254.
[93] Han Zhongyi, Zhou Guanglin, He Rundong, et al. How Well Does GPT-4V(ision) Adapt to Distribution Shifts? A Preliminary Investigation. arXiv preprint arXiv: 2312.07424, 2023.
[94] Song Han, Jeff Pool, John Tran, et al. Learning both weights and connections for efficient neural network. Advances in neural information processing systems, 2015, 28.
[95] Deng Lei, Li Guoqi, Han Song, et al. Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE, 2020, 108(4): 485-532.
[96] Gou Jianping, Yu Baosheng, Stephen John Maybank, et al. Knowledge distillation: A survey. International Journal of Computer Vision, 2021, 129: 1789-1819.