Generalized Video Anomaly Event Detection: Systematic Taxonomy and Comparison of Deep Models

YANG LIU, Academy for Engineering & Technology, Fudan University, China
DINGKANG YANG, Academy for Engineering & Technology, Fudan University, China
YAN WANG, Academy for Engineering & Technology, Fudan University, China
JING LIU, Academy for Engineering & Technology, Fudan University, China
LIANG SONG*, Academy for Engineering & Technology, Fudan University, China

Video Anomaly Event Detection (VAED) is the core technology of intelligent surveillance systems aiming to temporally or spatially locate anomalous events in videos. With the penetration of deep learning, the recent advances in VAED have diverged various routes and achieved significant success. However, most existing reviews focus on traditional and unsupervised VAED methods, lacking attention to emerging weakly-supervised and fully-unsupervised routes. Therefore, this review extends the narrow VAED concept from unsupervised video anomaly detection to Generalized Video Anomaly Event Detection (GVAED), which provides a comprehensive survey that integrates recent works based on different assumptions and learning frameworks into an intuitive taxonomy and coordinates unsupervised, weakly-supervised, fully-unsupervised, and supervised VAED routes. To facilitate future researchers, this review collates and releases research resources such as datasets, available codes, programming tools, and literature. Moreover, this review quantitatively compares the model performance and analyzes the research challenges and possible trends for future work.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Applied computing** → *Surveillance mechanisms*; • **Information systems** → Data streaming.

Additional Key Words and Phrases: Video anomaly event detection, intelligent video surveillance system, deep learning; normality learning, multiple instance learning

ACM Reference Format:
Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, and Liang Song. 2023. Generalized Video Anomaly Event Detection: Systematic Taxonomy and Comparison of Deep Models. 1, 1 (February 2023), 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Surveillance cameras can sense environmental spatial-temporal information without contact and have been the primary data collection tool for public services such as security protection [25], crime warning [144], and traffic management [90]. However, with the rapid development of smart cities and digital society, the number of surveillance cameras is

* Corresponding author.

Authors' addresses: Yang Liu, yang_liu20@fudan.edu.cn, Academy for Engineering & Technology, Fudan University, Shanghai, China, 200433; Dingkang Yang, dkyang20@fudan.edu.cn, Academy for Engineering & Technology, Fudan University, Shanghai, China, 200433; Yan Wang, yanwang19@fudan.edu.cn, Academy for Engineering & Technology, Fudan University, Shanghai, China, 200433; Jing Liu, jingliu19@fudan.edu.cn, Academy for Engineering & Technology, Fudan University, Shanghai, China, 200433; Liang Song, songl@fudan.edu.cn, Academy for Engineering & Technology, Fudan University, Shanghai, China, 200433.

Manuscript submitted to ACM

growing explosively, making the ensuing video analysis a significant challenge. Traditional manual inspection is time-consuming and laborious and may cause missing detections due to human visual fatigue [89], hardly coping with the vast scale video stream. As the core technology of intelligent surveillance systems, Video Anomaly Event Detection (VAED) aims to automatically analyze video patterns and locate abnormal events. Due to its potential application in unmanned factories, self-driving vehicles, and secure communities, VAED has received wide attention from academia and industry.

VAED in a narrow sense refers specifically to Unsupervised Video Anomaly Detection (UVAD), a paradigm that uses only normal videos to learn a normality model [124]. UVAD shares the same assumption as the long-established Anomaly Detection (AD) tasks in non-visual data (e.g., time series [7] and graphs [104]). They assume the normality model learned on normal samples cannot represent anomalous samples. Typically, UVAD consists of two phases, normality learning, and downstream anomaly detection [46, 84, 90]. UAVD shares a similar modeling process with other AD tasks without predefining and collecting anomalies, following the open-world principle. In the real world, anomalies are diverse and rare, so they cannot be defined and fully collected in advance. Therefore, UVAD was favored by early researchers and was once considered the prevailing VAED paradigm. However, the definition of anomaly is idealistic, ignoring that normal events are diverse. It is also unrealistic to collect all possible regular events for modeling. In addition, the learned UVAD model has difficulty maintaining a reasonable balance between representation and generalization power, either due to the insufficient representational that false-alarms unseen normal events as anomalies or the excessive generalization power that effectively reconstructs anomalous events. Numerous experiments [192] have shown that UVAD is valid for only simple scenarios. The model performance on complex datasets [84] is much inferior to that of simple single-scene videos [78, 94], which limits the application of VAED in realistic scenarios.

In contrast, Weakly-supervised Abnormal Event Detection (WAED) departs from the ambiguous setting that all are anomalous except normal with a clearer definition for the anomaly that is more consistent with human consciousness (e.g., traffic accidents, robbery, stealing, and shooting). [144]. Given its potential for immediate references in real-life applications such as traffic management platforms and violencFUVADe warning systems, WAED has become another mainstream VAED technical route [35, 88, 149]. Generally, WAED models directly output anomaly scores by comparing the spatial-temporal features of normal and abnormal events through Multiple Instance Learning (MIL). The previous study [35] proved that WAED could understand the essential difference between normal and abnormal. Therefore, its results are more reliable than that of UVAD. Unfortunately, WAED does not follow the basic assumptions of AD tasks. It is more like a binary classification under unique settings (e.g., data imbalance, negative samples containing multiple subcategories). Therefore, existing reviews [12, 121, 124] mainly focus on UVAD and consider WAED as a marginal route, lacking the organization and classification for WAED datasets and methods.

As stated, UVAD requires the training set to contain only normal videos, which is essentially semi-supervised learning in an extreme setting (training samples are all labeled as 0). To directly utilize a large number of raw videos without preemptive manual filtering, researchers [116, 182] proposed Fully-unsupervised Video Anomaly Detection (FVAD). FVAD needs no labels and does not impose any restrictions on the training data, learning the anomaly detector directly from the unscreened videos. The concept of UVAD indicates modeling with only normal data is well entrenched, so we still refer to such methods as UVAD and call the emerging truly unsupervised route as FVAD to avoid ambiguity.

In summary, this review focuses on anomaly event detection in surveillance videos, integrating deep VAED methods based on different assumptions, learning paradigms, and supervision into a systematic taxonomy: Generalized Video Anomaly Event Detection (GVAED), as shown in Fig. 1. We compare the differences and performance among different methods, sorting out the recent advances in GVAED. In addition, we collate available research resources, such as

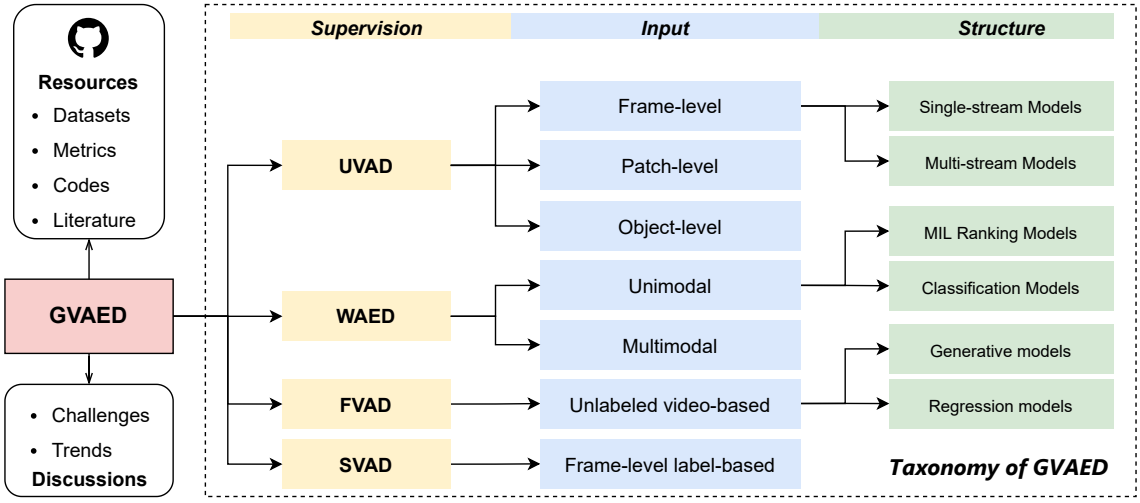

Fig. 1. Taxonomy of Generalized Video Anomaly Event Detection (GVAED). We provide a hierarchical taxonomy that organizes existing deep GVAED models by supervised signals, model inputs, and network structure into a systematic framework, including Unsupervised Video Anomaly Detection (UVAD), Weakly-supervised Abnormal Event Detection (WAED), Fully-unsupervised VAD (FVAD) and Supervised VAD (SVAD). Besides, we collate benchmark datasets, evaluation metrics, available codes, and literature to a public GitHub repository[1]. Finally, we analyze the research challenges and possible trends.

datasets, metrics, codes, and literature, into a public GitHub repository[1]. Moreover, we analyze the research challenges and future trends, which can guide further research and promote the development and applications of GVAED.

1.1 Literature Statistics

We count the publications and citations of academic papers on the topic of *Video Anomaly Detection* and *Abnormal Event Detection* in the past 12 years through reference databases (e.g., ACM Digital Library, IEEE explore, ScienceDirect, and SpringerLink) and search engines. The results are shown in Fig. 2, where the bar and dashboard represent the number of publications and citations. The dashes in Fig. 2(a) and Fig. 2(b) show a steadily increasing trend, indicating that GVAED has received wide attention. Therefore, a systematic taxonomy and comparison of GVAED methods are necessary to guide further development. As mentioned above, current works focus on unsupervised methods that use only regular videos to train models to represent normality. Thus, the development of UVAD is limited by representation means. Until 2016, UVAD utilized handicraft features, such as Local Binary Patterns (LBP) [53, 110, 190], Histogram Of Gradients (HOG) [46, 103, 133], and Space-Time Interest Point (STIP) [30]. The performance is poor and relies on a priori knowledge. As a result, VAED developed slowly. Fortunately, after 2016, with the development of deep learning, especially the application of Convolutional Neural Networks (CNNs) in image processing [19] and video understanding [184], VAED has ushered in new development opportunities. The research progress increased significantly, as shown in Fig 2(a). Deep CNNs can extract the video patterns end-to-end, freeing VAED research from complex a priori knowledge construction. In addition, compared to manual features [30, 53, 133], deep representations can capture multi-scale spatial semantic features and extract long-range temporal contextual features, which are more efficient in learning video normality. On the one hand, the large amount of video generated by the surveillance cameras provides sufficient training data for deep GVAED models. On the other hand, the iteratively updated Graphics Processing Units (GPUs)

[1] https://github.com/fudanyliu/GVAED.git

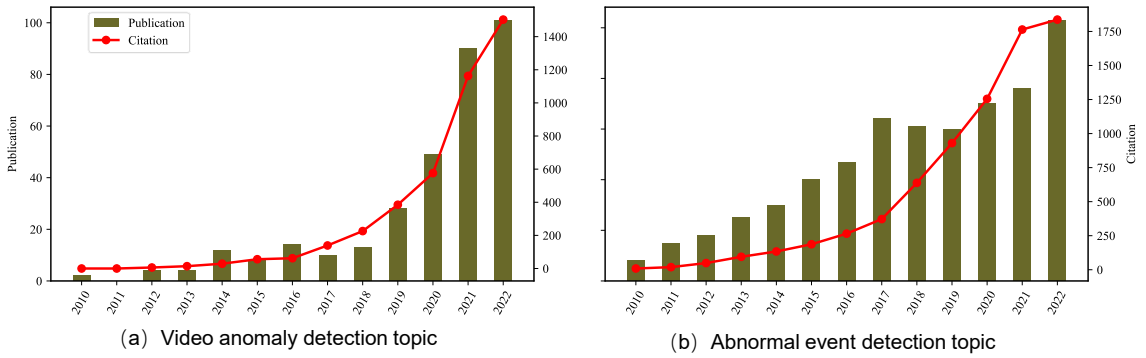

Fig. 2. Publication (Bar) and citation (Line) statistics on the topic of (a) *Video Anomaly Detection* and (b) *Abnormal Event Detection*.

make it possible to train large-scale models. As a result, VAED has developed rapidly in recent years and started to move from academic research to commercial applications. Similarly, Fig 2(b) reflects the research enthusiasm and development potential of abnormal event detection. To accelerate the application of GVAED in terminal devices and inspire future researchers, this review organizes various GVAED models into a unified framework. Additionally, we collect commonly used datasets, publicly available codes, and classic literature for further research.

1.2 Related Reviews

In the past four years, several reviews [7, 12, 23, 59, 61, 93, 109, 112, 115, 121, 124, 129, 135, 139] have covered GVAED works and generated various classification systems. Since GVAED is derived from the cross-pollination of video understanding and anomaly detection, some reviews [7, 124] that do not object to video anomaly detection also mention GVAED models. We analyze the methodologies covered in recent reviews and the research topics related to real-world deployment, as shown in Table 1. The mainstream reviews [61, 124] still consider VAED as a narrowly unsupervised task, lacking attention to WAED with excellent application value and FVAD methods using unlimited training data. In addition, they are biased against Supervised Video Anomaly Detection (SVAD), arguing that data labeling makes SVAD challenging to develop. However, the game engines [34, 137] and automatic annotations [38, 45] make it possible to obtain anomalous events and fine-grained labels. In addition, the existing review suffers from three major weaknesses: (1) [124] and [135] attempted to link existing works to specific scenes, missing the cross-scene challenges in real-world. Specifically, [124] pointed out that existing works were trained and tested on videos of the same scene, so they only reviewed single-single methods, leaving out the latest crass-scene VAED research. [135] focuses on the traffic VAED methods, innovatively analyzing the applicability of existing works in traffic scenes. However, weakly-supervised methods for crime and violence detection fail to be included in [135]. (2) Due to timeliness, earlier reviews [23, 112, 129, 135] were unable to cover the latest research and were outdated for predicting research trends. Recent surveys [12, 121] lack discuss the interaction of GVAED with new techniques such as causal inference [82, 173], domain adaptation/generalization [40, 165, 196], diffusion models [24], and online evolutive learning [70, 71, 142], which are expected to be the future directions of GVAED and essential to model deployment. (3) Although the latest review [12] in 2022 has started to incorporate WAED and SVAD, it still treats them as a marginal exploration, lacking a systematic organization of the datasets, literature, and trends.

Table 1. Analysis and Comparison of Related Reviews.

Year	Ref.	Main Focus	Methods[a]				Topics[a,b]			
			UVAD	WAED	SVAD	FVAD	LW	OD	CS	OE
2018	[61]	Unsupervised and semi-supervised methods.	✓	✗	✗	✗	✓	✗	✗	✗
2019	[93]	Weakly-supervised VAED methods and applications.	✗	✓	✗	✗	✗	✗	✗	✗
2020	[124]	Unsupervised sinle-scene video anomaly detection.	✓	✗	✓	✗	✓	✗	✓	✗
2021	[112]	Deep learning driven unsupervised VAED methods.	✓	✗	✗	✗	✓	✗	✗	✗
2021	[129]	Unsupervised crowd anomaly detection methods.	✓	✗	✗	✗	✗	✗	✗	✗
2021	[135]	Traffic scene video anomaly detection.	✓	✗	✗	✗	✗	✗	✗	✗
2022	[121]	Unsupervised video anomaly detection	✓	✗	✗	✗	✗	✓	✗	✗
2022	[12]	One&two-calss classification-based methods.	✓	✓	✓	✗	✓	✓	✗	✗
2023	Ours	GVAED taxonomy, challenges and trends.	✓	✓	✓	✓	✓	✓	✓	✓

a: ✗means no systematic analysis, while ✓is vice versa. b: LW=lightweight, OD=online detection, CS=cross-scene, and OE=online evolution.

1.3 Contribution Summary

GVAED will usher in new development opportunities with the rapid growth of deep learning technicals and surveillance videos. To clarify the development of GVAED and inspire future research, this review integrates UVAD, WAED, FVAD, and SVAD into a unified framework from an application perspective. The main contributions of this review are summarized in the following four aspects:

(1) To the best of our knowledge, it is the first comprehensive review that extends video anomaly detection from narrowly unsupervised methods to generalized video anomaly event detection. We analyze the various research routes and clearly state the lineage and trends of deep GVAED models to help advance the field.
(2) We organize various GVAED methods with different assumptions and learning frameworks from an application perspective, providing an intuitive taxonomy based on supervision signals, input data, and network structures.
(3) This review collects accessible datasets, literature, and codes and makes them publicly available. Moreover, we analyze the potential applications of other deep learning techniques and structures in GVAED tasks.
(4) We look at the research challenges and trends of GVAED in the context of deep learning development and intelligent video surveillance system deployment, which are expected to guide future researchers and engineers.

The remainder of the paper is organized as follows. Section 2 provides an overview of the basics and research background of GVAED, including the definition of anomalies, basic assumptions, main evaluation metrics, and benchmark datasets. Sections 3-6 introduce the unsupervised, weakly-supervised, fully-unsupervised, and supervised GVAED methods. We analyze the extant methods' general ideas and specific implementations and compare their strengths and weaknesses. Further, we quantitatively compare the performance in Section 6. Section 7 analyzes the challenges and research outlook on the development of GVAED. Section 8 concludes this review.

2 FOUNDATIONS OF GVAED

2.1 Anomaly Definition

UVAD follows the assumption of general AD tasks [115] and considers all events that have not occurred in the training set as abnormal. In other words, the training set of the UVAD dataset contains only normal events, while videos in the test set that differ from the training set are considered anomalies. Thus, certain normal events in the subjective human consciousness may also be labeled anomalies. For instance, in the UCSD Pedestrian datasets [78], riding a bicycle on the college campus is labeled abnormal simply because the training set fails to contain such events. This seemingly odd

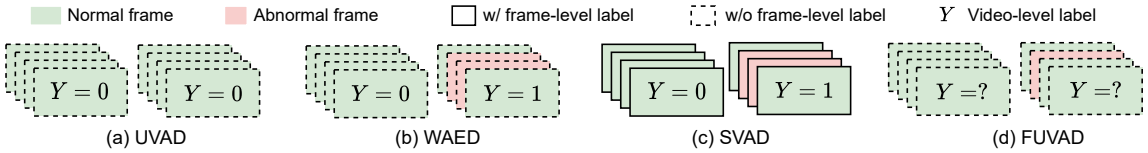

Fig. 3. Illustration of training data. (a) UVAD trains the model using only normal data, with the hidden implication that all video-level and frame-level labels are 0. (b) WAED models use positive and negative samples and require frame-level labels, where $Y = 0$ indicates normal video and $Y = 1$ indicates an anomaly. (c) SVAD is trained using a fine-grained frame-level labeling supervised model, where the semantics of the frame-level labels expose the video-level labels. (d) FVAD attempts to learn the anomaly detector from under-processed data with training data containing both normal and anomalous samples and without any level of labeling.

definition is dictated by the diversity and rarity of real-world anomalies. Collecting a sufficient number of anomalous events with a full range of categories is nearly impossible. In response, researchers have taken the alternative route of collecting enough normal videos to train models to describe the boundary of normal patterns and treat events that fall outside the boundary as anomalies. Unfortunately, it is also costly to collect all possible normal events for training. In addition, abnormal and normal frames share most of the appearance and motion information, making their patterns overlap. Therefore, letting the model find a discriminative pattern boundary without seeing abnormal events is infeasible.

In contrast, WAED takes a more intuitive definition of anomalies. Events that are subjectively perceived as abnormal by humans are considered anomalies, such as thefts and traffic accidents [144]. The training set for WAED tasks contains both normal and abnormal events and provides easily accessible video-level labels to supervise the model. Compared with fine-grained frame-level labels, video-level labels only tell the model whether a video contains abnormal events without revealing the exact location of the abnormalities, avoiding the costly frame-by-frame labeling and providing more reliable supervision. In contrast, the discrete frame-level annotations (0=normal, 1=abnormal) in SVAD ignore the transition continuity from normal to abnormal events. WAED needs to predefined abnormal events so that it can only distinguish specified abnormal events.

2.2 Problem Formation

The traditional unsupervised approaches [46, 84] treat GVAED as an outlier detection task, i.e., outlier events that are distinct from normal events are all anomalous. In UVAD, the training data contains only normal events, as shown in Fig. 3(a). The UVAD methods aim to describe the boundaries of normal spatial-temporal patterns with a proxy task and consider the test samples whose patterns fall outside the learned boundaries as anomalies. Thus, the degree of abnormality is usually quantified based on the deviation of the test sample from the learned model, with the logic that a model learned on massive normal events cannot effectively represent unseen abnormal events. Fig. 4 shows the two-stage anomaly detection framework in UVAD. The deep network trained by performing the proxy task in the training phase is directly applied as a normality model for anomaly detection in the testing phase. The performance of the proxy task is a credential to calculate the anomaly score. Formulaically, the process of UVAD is as follows:

$$e = d(f(\boldsymbol{x}_{\text{test}}; \theta), \boldsymbol{x}_{\text{test}}) \tag{1}$$

where θ denotes the learnable parameters of the deep model f, designed to characterize the prototype of normal events. d denotes the deviation between the test sample x_{test} and the well-trained f, which is usually a quantifiable distance, such as the Mean Square Error (MSE) of the prediction result, the L_2 distance in the feature space and the difference

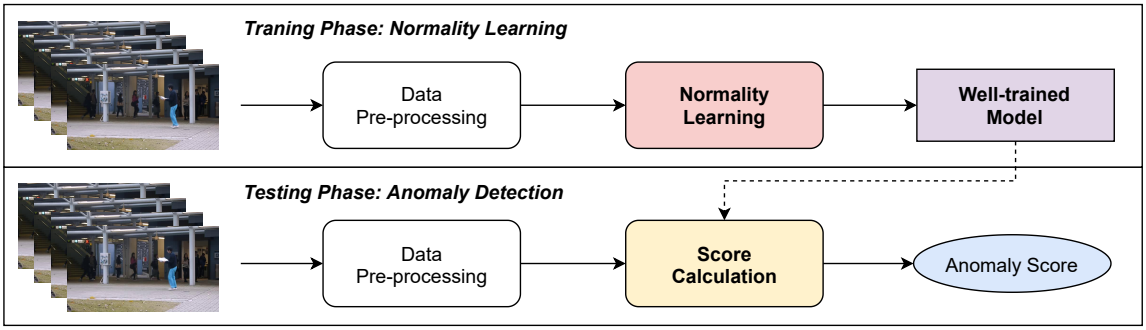

Fig. 4. Illustration of the two-stage UVAD framework. Anomaly detection is performed in the test phase as a downstream task of proxy task-based normality learning. The example video frames are from the CUHK Avenue [94] dataset.

of the distribution [100, 118, 193]. Noting that the normality model is obtained by optimizing the proxy task. This process is independent of the downstream anomaly detection, so the performance of the proxy task cannot directly determine the anomaly detection performance. In addition, for the reconstruction-based [42, 46] and prediction-based [84, 90] methods, the final anomaly score is usually a relative value in the range $[0, 1]$. A higher score indicates a larger deviation. Generally, these methods convert the absolute deviation e into a relative anomaly score by performing maximum-minimum normalization. They not only explicitly require all training data to be normal but also include the hidden assumption that the test videos must include anomalous events. In other words, any test video will yield high score intervals, which indicates that such methods are offline and may produce false alarms for normal videos.

WAED methods [35, 88, 149] always follow the MIL ranking framework [144]. A video is regarded as a collection of instances (segments), and whether it contains anomalies is known, while the location of the abnormal frames is unknown. Fig. 3(b) shows the training data composition for WAED, where both normal and anomalous events need to be pre-collected and labeled. Video-level labels are easy to obtain and often more accurate than the fine-grained frame-level labels for SVAD shown in Fig. 3(c). In a concrete implementation, WAED treats the video as a bag containing several instances, as illustrated in Fig. 5. The normal video $\mathcal{V}_n$ forms a negative bag $\mathcal{B}_n$, while the abnormal video $\mathcal{V}_a$ a positive bag $\mathcal{B}_a$. Based on MIL, WVEAD aims to train a regression model $r(\cdot)$ to assign scores to instances, with the basic goal that the maximum score of $\mathcal{B}_a$ is higher than that of $\mathcal{B}_n$. Thus, the WAED methods do not rely on an additional self-supervised proxy task but compute anomaly scores directly. The objective function $O(\mathcal{B}_a, \mathcal{B}_n)$ is as follows:

$$O(\mathcal{B}_a, \mathcal{B}_n) = \min \max \left(0, 1 - \max_{i \in \mathcal{B}_a} r\left(\mathcal{V}_a^i\right) + \max_{i \in \mathcal{B}_n} r\left(\mathcal{V}_n^i\right)\right) + \lambda_1 \overbrace{\sum_i^{n-1} \left(r\left(\mathcal{V}_a^i\right) - r\left(\mathcal{V}_a^{i+1}\right)\right)^2}^{C_{sparsity}} + \lambda_2 \overbrace{\sum_i^{n} r\left(\mathcal{V}_a^i\right)}^{C_{smooth}} \quad (2)$$

In addition to the additional anomaly curve smoothness constraint C_{sooth} and sparsity constraint $C_{sparsity}$, the core of $O(\mathcal{B}_a, \mathcal{B}_n)$ is to train a ranking model capable of distinguishing the spatial-temporal patterns between $\mathcal{B}_a$ and $\mathcal{B}_n$. Subsequent WAED works [35, 88, 91, 149, 157, 198] have followed the idea of MIL ranking and made effective improvements regarding feature extraction [198], label denoising [194], and the objective function [88]. However, as shown in Fig. 5, the MIL regression module take the extracted feature representations as input, so the performance of WAED methods partially depends on the pre-trained feature extractor, making the calculation costly.

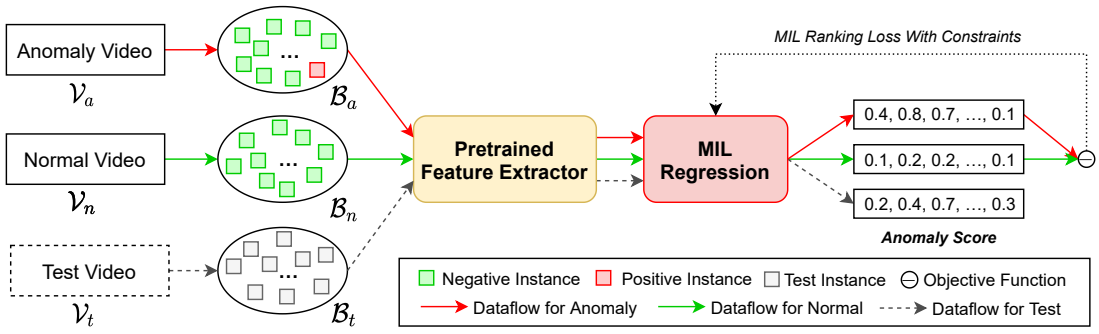

Fig. 5. Structure of the MIL ranking model [144]. The anomalous video $\mathcal{V}_a$ and the normal video $\mathcal{V}_n$ are first sliced into several equal-size instances. The positive bag $\mathcal{B}_a$ contains at least one positive instance, while the negative bag $\mathcal{B}_n$ contains only normal instances. In the test phase, the well-trained MIL regression model output the anomaly scores of instances in the test video $\mathcal{V}_t$ directly.

To completely avoid labeling costs and discard restrictions on training data, FVAD aims to use raw video to train the model. The training data contains both normal and abnormal, and none of the data labels are available for model training, as shown in Fig. 3(d). One class of FVAD methods follows a similar workflow to that of UVAD, i.e., learning the normality model directly from the original data. Although the training data contains anomalous events, the low frequency of anomalies limits their impact on model optimization. As a result, the model learned on many normal videos and a small number of abnormal frames is still only effective in representing normal events and generates large errors for anomalous events. Another class of methods tries to discover anomalies through the mutual collaboration of the representation learner and anomaly detector. Generally, the learning process of FVAD can be formulated as follows:

$$\mathcal{F} = \arg\min_{\Theta} \sum_{I \in \mathbb{I}} \mathcal{L}_{foc}(\hat{y} = \phi(m = \varphi(x = f(I))), l) \tag{3}$$

where the aim is to learn an anomaly detector $\mathcal{F}$ via a deep neural network which consists of a backbone network $f(\cdot;\Theta_b) : \mathbb{R}^{H\times W\times 3} \mapsto \mathbb{R}^{D_b}$ that transforms an input video frame I to feature $\boldsymbol{x}$, an anomaly representation learner $\varphi(\cdot;\Theta_a) : \mathbb{R}^{D_b} \mapsto \mathbb{R}^{D_n}$ that converts x to an anomaly specific representation m, and an anomaly score regression layer $\phi(\cdot;\Theta_s) : \mathbb{R}^{D_s} \mapsto \mathbb{R}$ that learns to predict m to an anomaly score y. The overall parameters $\Theta = \{\Theta_b, \Theta_a, \Theta_s\}$ are optimized by the focal loss.

Research on fully-unsupervised methods is still in its infancy, and they exploited the imbalance of samples and the significant difference of anomalies in the GVAED task.

2.3 Benchmark Datasets

Public datasets provide a fair benchmark for model testing. Their development also reflects the trends and research fervor. In Table 2, we show and compare the statistical results and properties of the frequently used GVAED datasets. Several datasets [1, 84, 144, 168] have been proposed with different annotated signals to match new research requirements after 2018, which reflects the trend of GVAED from unsupervised to weakly-unsupervised [144], from unimodal to multimodal [168] and from simple to complex real-world scenarios [84] at the data level.

2.3.1 Subway Entrance & Exit. As an earlier dataset, Subway [2] includes two independent sub-datasets, Entrance and Exit, which record the subway entrance and exit scenes, respectively. The anomalous events include people who skip

Table 2. Representative GVAED Datasets. *Italicized ones* indicate WAED datasets, and underlined one is multimodal dataset.

Year	Dataset	#Videos			#Frames					#Scenes	#Anomalies
		Total	Training	Testing	Total	Training	Testing	Normal	Abnormal		
2008	Subway Entrance[2]	-	-	-	144,250	76,543	67,797	132,138	12,112	1	51
2008	Subway Exit[2]	-	-	-	64,901	22,500	42,401	60,410	4,491	1	14
2011	UMN[3†]	-	-	-	7,741	-	-	6,165	1,576	3	11
2013	UCSD Ped1[4]	70	34	36	14,000	6,800	7,200	9,995	4,005	1	61
2013	UCSD Ped2[4]	28	16	12	4,560	2,550	2,010	2,924	1,636	1	21
2013	CUHK Avenue[5]	37	16	21	30,652	15,328	15,324	26,832	3,820	1	77
2018	ShanghaiTech[6]	-	-	-	317,398	274,515	42,883	300,308	17,090	13	158
2018	*UCF-Crime*[7]	1,900	1,610	290	13,741,393	12,631,211	1,110,182	-	-	-	950
2019	*ShanghaiTech Weakly*[8]	437	330	107	-	-	-	-	-	-	-
2020	Street Scene[9]	81	46	35	203,257	56,847	146,410	159,341	43,916	205	17
2020	*XD-Violance*[10]	4,754	-	-	-	-	-	-	-	-	-
2022	UBnormal[11] ‡	543	268	211	236,902	116,087	92,640	147,887	89,015	29	660

† Following previous works, we set the frame rate to 15 fps. ‡ The UBnormal dataset is supervised and includes a validation set with 64 videos.

the subway entrance to evade tickets, cleaners who behave differently from regular entry and exit, and people who travel in the wrong direction. Due to the cursory nature of the labeling work and the lack of clarity in the definition of anomalous events, most existing works have refrained from using this dataset for model evaluation. Therefore, we do not provide quantitative performance comparison results on this dataset but only briefly describe its characteristics to reflect the lineage of GVAED datasets development.

2.3.2 UMN. The UMN [25] is also an early GVAED dataset containing 11 short videos captured from three different scenes: grassland, indoor hall, and park. The scenes are set by the researcher rather than naturally filmed to detect abnormal crowd behavior in indoor and outdoor scenes, i.e., the crowd suddenly shifts from normal interaction to evacuation and fleeing abruptly to simulate fear. The anomalies are artificially conceived and played out, ignoring the diversity and rarity of anomalies in the real-world. Similar to the Subway [2] dataset, UMN has been abandoned by recent researchers due to the lack of spatial annotation.

2.3.3 UCSD Pedestrian. UCSD Ped1 & Ped2 [78] are the most widely used UVAD datasets collected from university campuses with simple but realistic scenarios. They reflected the value of GVAED in public security applications. Specifically, the Ped1 dataset is captured by a camera with a viewpoint perpendicular to the road, so the moving object's size changes with its spatial position. In contrast, the Ped2 dataset used a camera whose viewpoint is parallel to the direction of the road, which is simpler than Ped1. Pedestrian walking is defined as normal, while behaviors and objects different from it are considered abnormal, such as biking, skateboarding, and driving. Since the scene is classical and anomalous events are easy to understand, UCSD Pedestrian is widely used by existing works, and the frame-level AUC has been as high as 99%, reflecting the saturation of model performance. Therefore, the dataset in simple scenes has become a constraint for GVAED development. Large-scale, cross-scene, multimodal complex datasets have become an inevitable trend.

2.3.4 CUHK Avenue. Similar to UCSD Pedestrian, the CUHK Avenue [94] dataset is also collected from the university campus, and both focus on anomalous events that occur on the road outside of expectations. The difference is that most of

[2] https://vision.eecs.yorku.ca/research/anomalous-behaviour-data/sets/ [3] http://mha.cs.umn.edu/proj_events.shtml#crowd
[4] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm [5] http://www.cse.cuhk.edu.hk/leojia/projects/detectabnormal/dataset.html
[6] https://svip-lab.github.io/dataset/campus_dataset.html [7] https://webpages.charlotte.edu/cchen62/dataset.html
[8] https://github.com/jx-zhong-for-academic-purpose/GCN-Anomaly-Detection/ [9] https://www.merl.com/demos/video-anomaly-detection
[10] https://roc-ng.github.io/XD-Violence/ [11] https://github.com/lilygeorgescu/UBnormal

Fig. 6. Examples of XD-Violence dataset [168]. XD-Violence is a multimodal dataset for violence detection, including video and audio. We show video frames here. The anomalous events are not all from the real-world but also include movie and game footage, etc.

the 47 anomalous events in CUHK Avenue are simulated by the data collector, including appearance anomalies (e.g., bags placed on the grass) and motion anomalies, such as throwing and wrong direction. CUHK Avenue provides both frame-level and pixel-level spatial annotations. In addition, its large data scale makes it one of the mainstream UVAD datasets.

2.3.5 ShanghaiTech. Considering that the UCSD Pedestrian [78] and CUHK Avenue [94] datasets only consider anomalous events in a single scene, while the real world usually faces the challenge of spatial-temporal pattern shifts across scenes, for this reason, the team from ShanghaiTech University proposed the ShanghaiTech [84] dataset containing 13 scenes, providing the largest UVAD benchmark. Abnormal behaviors are defined as all collected behaviors that distinguish them from normal walking, such as riding a bicycle, crossing a road, and jumping forward. Unfortunately, although the collectors pointed out the shortcomings of the existing dataset with a single scenario, their proposed FFP [84] was not explicitly designed to address the cross-scene challenges but rather to treat it as a whole without differentiating between scenarios. For the WAED setting, researchers [194] proposed to move some anomalous videos from the test set to the training set and provided video-level labels for each training video, introducing the ShanghaiTech Weakly dataset, which has become the mainstream WAED benchmark. An interesting phenomenon is that the performance of WAED methods on ShanghaiTech Weakly (frame-level AUC is typically > 85% and has reached up to 95%) is generally higher than that of UVAD methods on the ShanghaiTech (frame-level AUC is typically between 70 ∼ 80%), providing evidence for the applicability of WAED in complex scenarios over UVAD.

2.3.6 UCF-Crime. UCF-Crime [144] is the first WAED dataset, presented together with the original MIL ranking framework. UCF-Crime consists of 1900 unedited real-world surveillance videos collected from the Internet. Anomalous events are objected to people's concerns, such as burglaries, robberies, and traffic accidents, with a total of 850 anomalous events classified into 13 categories. Unlike the UVAD dataset above, the UCF-Crime's training set contains anomalous videos and provides a video-level label for each video, where 0 indicates normal, and 1 indicates anomalous. The anomalous events in the WAED dataset are predefined and are usually associated with specific scenarios, such as car accidents in urban traffic, shoplifting, and shootings in neighborhoods. Therefore, WAED can provide more credible results for real scenarios with better application potential.

2.3.7 XD-Violence. As the first audio-video dataset, XD-Violence [168] expands anomaly event detection from single-modal video understanding to multimodal signal processing, facilitating the coexistence of GVAED and multimedia communities. XD-Violence focuses on violent behaviors, such as abuse, explosion, car accident, struggle, shootings,

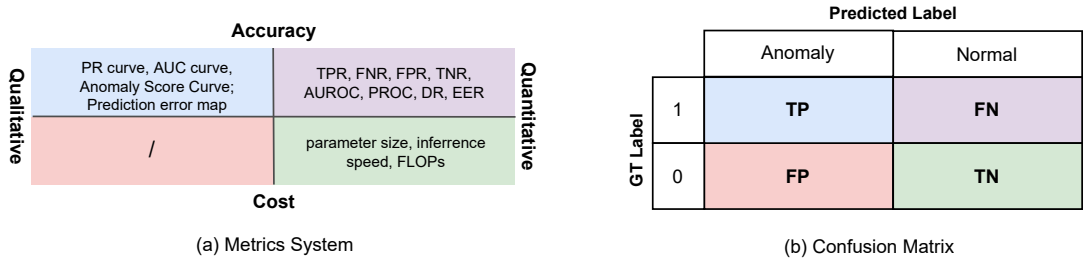

Fig. 7. Illustration of GVAED performance evaluation system. We show the (a) metrics system and (b) confusion matrix.

and riots, as shown in Fig. 6. Due to the rarity of violent behaviors and the high difficulty of capturing violence, the original videos include some movie clips in addition to real-world surveillance videos. XD-Violence provides a new way to think about the GVAED by extending the data modality from single videos to sound, text, and others.

2.3.8 UBnormal. Inspired by the computer vision community benefiting from synthetic data, Acsintoae *et al.* [1] propose the first GVAED benchmark with virtual scenes, named UBnormal. Notably, Utilizing a data engine to synthesize data under predetermined instructions rather than collecting real-world data makes pixel-level labeling possible. Therefore, UBnormal is supervised. UBnormal is built to address the problem that WAED ignores the open-set nature of anomalies that prevents the model from correctly corresponding to new anomalies. The test set contains anomalous events not present in the training set. Moreover, it provides a validation set for model tuning for the first time.

2.4 Performance Evaluation

Existing GVAED methods evaluate model performance in terms of detection accuracy and operational cost. The former concerns the ability to discriminate anomalous events while the latter aims to measure the deployment potential on resource-limited devices. According to the scale of detected anomalies, the detection accuracy criteria are divided into three levels: Temporal-Detection-Oriented (TDO), Object-Detection-Oriented (ODO), and Spatial-Localization-Oriented (SLO). Specifically, TDO criteria require the model to determine anomalous events' starting and ending temporal position without spatial localization of abnormal pixels. In contrast, ODO criteria include object-level, region-level, and track-level, focusing on specific anomaly objects or trajectories. SLO criteria encourage pixel-level localization of anomalous events. As for operational cost criteria, the commonly used metric include parameter size, inference speed, and the number of floating point operations on the same platform, as shown in Fig. 7(a).

We can evaluate the model performance quantitatively by comparing the predicted results with the ground truth labels. It is worth noting that the predicted labels of some GVAED models (e.g., prediction-based UVAD and WVED) are continuous values in the range of $[0, 1]$. In contrast, the true labels are discrete 0 or 1, so a threshold value must first be selected when calculating the performance metrics. Samples with abnormal scores below the threshold are considered normal, and vice versa. In this way, we obtain the confusion matrix shown in Fig. 7(b), where TP, FN, FP, and TN denote the number of abnormal samples correctly detected, abnormal samples mistakenly detected as normal, normal samples mistakenly detected as abnormal, and normal samples correctly detected, respectively. The True-Positive-Rate (TPR), False-Positive-Rate (FPR), True-Negative-Rate (TNR), and False-Negative-Rate (FNR) are defined as follows:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}; \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}; \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{FP+TN}}; \mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{TP+FN}} \tag{4}$$

which are used to calculate the Area Under the Receiver Operating Characteristic (AUROC) and Average Precision (AP).

AUROC: The horizontal and vertical coordinates of the Receiver Operating Characteristic (ROC) curve are the FPR and TPR, and the curve is obtained by calculating the FPR and TPR under multiple sets of thresholds. The area of the region enclosed by the ROC curve and the horizontal axis is often used to evaluate binary classification tasks, denoted as AUROC. The value of AUROC is within the range of [0, 1], and higher values indicate better performance. AUROC can visualize the generalization performance of the GVAED model and help to select the best alarm threshold. In addition, the Equal Error Rate (EER), i.e., the proportion of incorrectly classified frames when TPR and FNR are equal, is also used to measure the performance of anomaly detection models.

AP: Due to the highly unbalanced nature of positive and negative samples in GVAED tasks, i.e., the TN is usually larger than the TP, researchers think that the area under the Precision-Recall (PR) curve is more suitable for evaluating GVAED models, denoted as AP. The horizontal coordinates of the PR curve are the Recall (R) shown in Eq. 4, while the vertical coordinate represents the Precision (P), defined as $P = \frac{TP}{TP+FP}$. A point on the PR curve corresponds to the P and R values at a certain threshold. Currently, AP has become the main metric for multimodal GVAED models [163, 168, 169] and is widely used to evaluate the performance on the XD-Violence dataset [168].

3 UNSUPERVISED VIDEO ANOMALY DETECTION

The UVAD method aims to learn a normality model using normal video and discriminate anomalies by measuring the deviation between the test sample and the learned model. Existing reviews [112, 124] usually classify UVAD methods into distance-based [26, 27, 133], probability-based [6, 20, 131, 134], and reconstruction-based [42, 46, 90] according to the deviations calculation means. Early traditional methods relied on manual features such as foreground masks [133], histogram of flow [27], motion magnitude [133], HOG [26], motion boundary histograms [63], dense trajectories [158], and STIP [30], which relied on human a priori knowledge and had poor representational power. With the rise of deep learning in computer vision tasks [136, 184, 191], recent approaches preferred to extracting features representations in an end-to-end framework with deep Auto-Encoders (AE) [21, 46, 87, 90, 118], Generative Adversarial Networks (GAN) [9, 17, 56, 69, 113, 188], and Vision Transformers (ViT) [36, 65, 179].

This section aims to provide a systematic overview of deep learning-driven UVAD methods. However, the traditional taxonomy [112, 124] focus on manual feature-based methods, which cannot help clarify the trends of deep UVAD. Therefore, we divide the works presented in the literature into three main groups depending on the type of data. Specifically, frame-level methods use whole RGB and flow frames as input and aims to learn the global normality. Considering the repetitive spatial-temporal information in the video sequence, patch-level methods extract features only from selected regions and ignore invalid information from repetitive regions as well as regional information interactions that do not require attention, which have advantages in terms of computational cost and inference speed. In recent years, researchers proposed to detect foreground objects and analyze the patterns of specific objects. The object-level methods consider the connection between the object and the background and perform excellently in the task of detecting anomalous events in complex scenes. Based on the above analysis, this review classifies the UVAD methods into frame-level, patch-level, and object-level according to the hierarchical *"input-structure"* taxonomy shown in Fig. 1.

3.1 Frame-level Methods

Deep CNNs can directly extract abstract features from videos and learn task-specific deep representations. Frame-level methods use complete RGB frames, sequences, or optical flows as input to model the normality of normal events in a

self-supervised learning manner. Existing methods can be classified into two categories according to model structure: single-stream and multi-stream. The former does not distinguish spatial and temporal information. They usually take the original RGB videos as input and learn the spatial-temporal patterns by reconstructing the input sequence or predicting the next frame. Existing single-stream work focuses on designing more efficient network structures. They introduced more powerful representational learners such as 3D convolution [193] and U-net [84]. In contrast, multi-stream networks typically treat appearance and motion as different dimensions of information and attempt to learn spatial and temporal normality using different agent tasks or network architectures. In addition to spatial-temporal separation modeling, existing dual-stream works explored spatial-temporal coherence [87] and consistency [8] to perform anomaly detection.

3.1.1 Single-Stream Models. Single-stream models typically use a single generative model to describe the spatial-temporal patterns of normal events by performing a proxy task and preserving the normality in learnable parameters. For example, the Predictive Convolutional Long Short-Term Memory (PC-LSTM) network in [108] used a conforming ConvLSTM network to model the evolution of video sequences. Hasan *et al.* [46] constructed a fully convolutional Feed-Forward Auto-Encoder (FF-AE) with manual features as input, which can learn task-specific representations in an end-to-end manner. Motivated by the representation ability of deep learning, Hu *et al.* [52] proposed a deep Incremental Slow Feature Analysis (D-IncSFA) network to complete feature extraction and anomaly detection together. Smeureanu *et al.* [141] used a pre-trained VGG network [143] to extract deep representation and performed anomaly detection using one-class SVM. ConvLSTM-AE [97] used CNN to encode frames and stored historical motion information by convolutional LSTM, which can better encode the appearance and motion and identify the abnormal variation.

Liu *et al.* [84] proposed a Future Frame Prediction (FFP) method that used a GAN-based video prediction framework to learn the normality. Its extension, FFPN [100], further specified the design principles of predictive UVAD networks. Singh and Pankajakshan [140] also used a predictive task to detect anomalies, proposing conformal structures based on 2D & 3D convolution and convLSTM to characterize spatial-temporal patterns more efficiently.

To address the detail loss in frame generation, Li *et al.* [79] proposed a Spatial-Temporal U-net network (STU-net) that combined the advantages of U-net in representing spatial information with the ability of convLSTM to model temporal variations for moving objects. [195] proposed a sparse coding-based neural network called AnomalyNet, which used three neural networks to integrate the advantages of feature learning, sparse representation, and dictionary learning. In [111], the authors proposed an Incremental Spatial-Temporal Learner (ISTL) to explore the nature of anomalous behavior over time. ISTL used active learning with fuzzy aggregation to continuously update and distinguish between new anomalous and normal events evolving. The anoPCN in [174] unified the reconstruction and prediction methods into a deep predictive coding network by introducing an error refinement module to reconstruct the prediction errors and refining the coarse predictions generated by the predictive coding module.

To lessen the deep model's ability to generalize anomalous samples, memory-augmented Auto-Encoder (memAE) [42] embedded an external memory network between the encoder and decoder to record the prototypical patterns of normal events. Further, Park *et al.* [118] introduced an attention-based memory addressing mechanism and proposed to update the memory pool during the testing phase to ensure that the network can better represent normal events.

Luo *et al.* [101] proposed a sparse coding-inspired neural network model, namely Temporally-coherent Sparse Coding (TSC). It used a sequential iterative soft thresholding algorithm to optimize the sparse coefficients. [28] introduces residual connection [49] into the auto-encoder to eliminate the gradient disappearance problem during normality learning. Experiments shown that ResNet brong 3%, 2% and 5% frame-level AUC gains for the proposed Residual Spatial-Temporal Auto-Encoder (R-STAE) on CUHK Avenue [94], LV [67] and UCSD Ped2 [78] datasets, respectively.

Table 3. Frame-level Multi-stream UVAD Methods.

Year	Method	Backbone	Analysis
2017	AMDN [171]	AE	**Pros:** Learning appearance and motion patterns separately. **Cons:** Determining boundaries by OC-SVM with limited capability.
2017	STAE [193]	AE	**Pros:** Using 3D CNN to learn the spatial-temporal patterns. **Cons:** Dual decoders causing huge computational costs.
2019	AMC [114]	GAN	**Pros:** Learning the correspondence between appearance and motion. **Cons:** Limited performance on the complex datasets.
2020	CDD-AE [14]	AE	**Pros:** Using two auto-encoders to learn spatial and temporal patterns. **Cons:** No special consideration in the design of the encoder structure.
2020	OGNet [180]	GAN	**Pros:** Using generators and discriminators to learn normality. **Cons:** Adversarial learning making the training process unstable.
2020	GANs [127]	GAN	**Pros:** Training two GANs to learn temporal and spatial distribution. **Cons:** Unstable traning process and high training cost.
2021	AMMC-net [8]	AE	**Pros:** Exploring the consistency of appearance and motion. **Cons:** Lack of analysis to the flow-frame generation task.
2021	DSTAE [76]	AE, ConvLSTM	**Pros:** Using two auto-encoders to perform different tasks. **Cons:** High computional cost and relying on optical flow network.
2022	AMAE [87]	AE	**Pros:** Using two encoders and three decoders to learn features. **Cons:** High training cost and relying on optical flow network.
2022	STM-AE [90]	AE, GAN	**Pros:** Using two memory-enhanced auto-encoders to learn normality. **Cons:** Unstable traning process and high computional costs.

The DD-GAN in [31] introduced an additional motion discriminator to GAN. The dual discriminators structure encouraged the generator to generate more realistic frames with motion continuity. Yu *et al.* [177] also used GAN to model normality. The proposed Adversarial Event Prediction (AEP) network performed adversarial learning on past and future events to explore the correlation. Similarly, Zhao *et al.* [192] explored spatial-temporal correlations by GAN and used a spatial-temporal LSTM to extract appearance and motion information within a unified unit.

[16] proposed a Bidirectional Prediction (Bi-Pre) framework that used forward and backward prediction sub-networks to reason about normal frames. In the test phase, only part significant regions are used to calculate the anomaly score, allowing the model to focus on the foreground. Wang *et al.* [162] used multi-path convGRU to perform frame prediction. The proposed ROADMAP model included three non-local modules to handle different scales of objects.

3.1.2 Multi-Stream Models. . The *multiplicity* of multi-stream models is reflected in the multiple sources of the input data and the multiple tasks corresponding to multiple outputs. Considering that video anomaly may manifest as outliers in appearance or motion, an intuitive idea is to use multi-stream networks to model spatial and temporal normality separately [13, 14, 76]. In addition, learning associations between appearance and motion, such as consistency[8], coherence [87, 90], and correspondence [114], is another effective GVAED solution. Events without such associations are discriminated against as anomalies. The multi-stream model has achieved significant success in recent years due to the high matching of its design motivation with the GVAED task. The multi-stream methods are summarized in Table 3.

Motivated by the remarkable success of 3D CNN in video understanding tasks, Zhao *et al.* [193] proposed a 3D convolutional-based Spatial-Temporal Auto-Encoder (STAE) to model normality by simultaneously performing reconstruction and prediction tasks. STAE included two decoders, which outputted reconstructed and predicted frames, respectively. In contrast, Appearance and Motion DeepNet (AMDN) [171] used two stacked denoising auto-encoders to encode RGB frames and optical flow separately. Similarly, Chang *et al.* [14] also used two auto-encoders to capture spatial and temporal information, respectively. One learned the appearance by reconstructing the last frame, while the other outputted RGB differences to simulate the generation of optical flow. Deep K-means clustering was used to force the extracted feature compact and detect anomalies. DSTAE [76] introduced convLSTM to a two-stream auto-encoder to better model the temporal variations. The reconstruction errors of the two encoders are weighted and used to calculate anomaly scores.

In addition to spatial-temporal separation, Nguyen *et al.* [114] proposed to learn the correspondence between appearance and motion. To this end, they proposed an AE with two decoders, one for reconstructing input frames and the other for predicting optical flow. Cai *et al.* [8] proposed an Appearance-Motion Memory Consistency network (AMMC-net), which aimed to capture the spatial-temporal consistency in high-level feature space.

Liu *et al.* [87] proposed an Appearance-Motion united Auto-Encoder (AMAE) framework using two independent auto-encoders to perform denoising and optical flow generation tasks separately. Moreover, they utilized an additional decoder to fuse spatial-temporal features and predict future frames to model spatial-temporal normality. STM-AE [90] introduced the memory into the dual-stream auto-encoder to record prototype appearance and motion patterns. Adversarial learning explores the connection between spatial and temporal information of regular events.

Aside from the above anomaly detection means such as reconstruction error [76, 87, 90, 193], clustering [13, 14] and one-class classification [171], researchers attempted to utilize the discriminator of GAN to directly output results. For instance, Ravanbakhsh *et al.* [127] used GAN to learn the normal distribution and detect anomalies directly by discriminators. The authors use a cross-channel approach to prevent the discriminator from learning mundane constant functions. OGNet [180] shifted the discriminator from discriminating real or generated frames to distinguishing good or poor reconstructions. The well-trained discriminator can find subtle distortions in the reconstruction results and detect non-obvious anomalies.

3.2 Patch-level Methods

The patch-level methods [86, 105, 130, 132] takes the video patch (spatial-temporal cube) as input. Compared with frame-level methods that consider anomalies roughly, i.e., anomalies are reflected in spatial or temporal dimensions beyond expectation, patch-level methods consider finding anomalies from specific spatial-temporal regions rather than analyzing the whole sequence. Patch formation can be divided into three categories: scale equipartition [22, 72, 86, 105, 132, 167, 197], information equipartition [68], and foreground object extraction [161]. Specifically, scale equipartition is the simplest. The video sequence is equipartitioned into several spatial-temporal cubes of uniform size along the spatial and temporal dimensions. The subsequent modeling process is similar to frame-level methods. The information equipartition strategy considers that image blocks of the same size do not contain the same information. Regions close to the camera contain less information per unit area than those far away. Before representation, all cubes will be first resized to the same size. The foreground object extraction focuses on modeling regions with information variation to avoid the learning cost and disruption of the background. After the sequences are equated into same-scale cubes, those containing only background will be eliminated.

Mehrsan *et al.* [130] densely sampled video sequences at different spatial and temporal scales and used a probabilistic framework to model the spatial-temporal composition of the video volumes. The STCNN [197] treated UVAD as a binary

Table 4. Patch-level UVAD Methods.

Year	Method	Patch Formulation	Detection Logic	Contributions
2010	ADCS [105]	Slicing the video into equal-sized spatial-temporal patches.	deviation to the learned mixtures of dynamic textures.	Detecting spatial and temporal anomalies by joint modeling of scene appearance and dynamics.
2013	STC [130]	Dense sampling at different spatial and temporal scales.	Modeling spatial-temporal arrangements of volumes.	Fast VAED model by coding spatial-temporal composition of volumes with a probabilistic framework.
2016	STCNN [197]	Equating the video sequence into $3 \times 3 \times 7$ pixel patches.	Binary classification with the fully connected network.	Detecting crowded scene anomalies by analyzing dynamic regions with spatial-temporal CNN.
2016	DeepAnomaly [22]	Slicing the video into equal-sized spatial-temporal patches.	Learning prototypical patterns of normal objects.	Detecting distant and occluded anomalies in agricultural scenes by combining CNN and background models.
2017	Deep-Cascade [132]	Equating the video sequence and resizing the object of interest.	Mahalanobis distance to the Gaussian models.	Cubic-patch-based time-efficient video anomaly localization method with a cascade of classifiers.
2017	sRNN [98]	Multiple Patches Sampled at Multiple Scales.	reconstruction error.	Mapping temporally-coherent Sparse Coding Using Stacked RNN.
2019	S^2-VAE [161]	Foreground extraction and keeping only the region of interest.	Reconstruction error.	Combining shallow stacked fully-connected variational AE and VAEs to detect local and global anomalies.
2020	DeepOC [167]	Equating the video frames and the corresponding optical flow.	One-class classification.	Using stacked convolutional encoders to generate low-dimensional features and training classifiers to make these features as compact as possible
2021	ST-CaAE [72]	Equating video sequences into spatial-temporal patch.	Reconstruction error.	First filter simple normal spatial-temporal patches and then describe normality by reconstruction strategies.
2022	AST-AE [86]	Equating video frames into 8×8 pathches along the spatial dimension.	Prediction error for some regions.	Using CNN and LSTM with channel attention mechanism to model spatial and temporal information, respectively.

classification task. It first extracted patches' appearance and motion information and outputted the discriminative results with an FCN. It first equated the video sequence into patches of $3 \times 3 \times 7$ and retained only the part of the region containing moving pixels to ensure the robustness of the model to local noise and improve the detection accuracy. Deep-Cascade [132] employed a cascaded autoencoder to represent video patches. It used a lightweight network to select local patches of interest and then applied a complex 3D convolutional network to detect anomalies. The lightweight network can filter simple normal patches to reduce computational costs and save processing time. S^2-VAE [161] first detected the foreground and retained only the cell containing the object as input. And then, a shadow generative network was used to fit the data distribution. The output was fed to another deep generative network to model normality. Wu *et al.* [167] proposed a deep one-class neural network (DeepOC). Specifically, DeepOC used stacked auto-encoders to generate low-dimensional features for frame and optical flow patches and simultaneously trained the OC classifier to make these representations more compact.

Table 5. Object-level UVAD Methods.

Year	Method	Detector	Decision Logic	Contributions
2017	LDGK [50]	Fast R-CNN	Anomaly score of the detected object proposal	Integrating a generic CNN and environment-related anomaly detector to detect video anomalies and record the cause of the anomalies.
2018	DCF [62]	YOLO	Classification	Extracting foreground objects by object detection models and Kalman filtering and discriminating anomalies by pose and motion classification.
2019	OC-AE [57]	SSD	One-versus-rest binary classification	Proposing an object-centric convolutional autoencoder to encode motion and appearance and discriminating anomalies using a one-versus-rest classifier.
2021	Background-Agnostic [40]	SSD-FPN, YOLOv3	Binary classification	Using a set of autoencoders to extract the appearance and motion features of foreground objects and then using a set of binary classifiers to detect anomalies.
2021	Multi-task [39]	YOLOv3	Binary classification	Training a 3D convolutional neural network to generate discriminative representation by performing multiple self-supervised learning tasks.
2021	OAD [33]	YOLOv3	Clustering	An online VAED method with asymptotic bounds on the false alarm rate, providing a procedure for selecting a proper decision threshold.
2021	HF2-VAD [92]	Cascade R-CNN	Prediction error	A hybrid framework that seamlessly integrates sequence reconstruction and frame prediction to handle video anomaly detection.
2020	VEC [176]	Cascade R-CNN	Cube construction error	Proposing a video event completion framework to exploit advanced semantic and temporal contextual information for video anomaly detection.
2022	BiP [15]	Cascade R-CNN	Appearance and motion error	Proposing a bi-directional architecture with three consistency constraints to regularize the prediction task from the pixel, cross pattern, and temporal levels.
2022	HSNBM [5]	Cascade R-CNN	Frame and object prediction error	Designing a hierarchical scene normative binding modeling framework to detect global and local anomalies.

Spatial-Temporal Cascade Auto-Encoder (ST-CaAE) [72] first used an adversarial autoencoder to identify anomalous videos and excluded simple regular patches. The retained patches were fed to a convolutional autoencoder, which discriminated anomalies based on reconstruction errors. Liu *et al.* [86] proposed an Attention augmented Spatial-Temporal Auto-Encoder (AST-AE) that equated frames in spatial dimensions into 8×8 parts and models spatial and temporal information using CNN and LSTM, respectively. In the downstream anomaly detection stage, AST-AE only retained significant regions with large prediction errors to calculate the anomaly score.

3.3 Object-level Methods

The emergence of high-performance object detection models [41, 128, 172] provides a new idea for GVAED, i.e., using a pre-trained object detector to extract the objects of interest from the video sequence before normality learning. Compared with the frame-level and patch-level methods, the object-level methods [40] enable the model to ignore redundant background information and focus on modeling the behavioral interactions of foreground objects. In addition

to outperforming object-free methods in terms of performance, object-level methods are also considered feasible to investigate scene-adaptive GVAED models. Existing studies [39, 40] show that object-level methods perform significantly better than other methods on multi-scene datasets such as ShanghaiTech [84]. Table 5 compares the object detectors, decision logic, and main contributions of existing object-level methods.

Ryota *et al.* [50] attempted to describe anomalous events in a human-understandable form by detecting and analyzing the classes, behaviors, and attributes of specific objects. The proposed LDGK model first used multi-task learning to obtain anomaly-related semantic information and then inserted an anomaly detector to analyze scene-independent features to detect anomalies. The DCF [62] used a pose classifier and an LSTM network to model the spatial and motion information of the detected objects, respectively. [57] formalizes UVAD as a one-versus-rest binary classification task. The proposed OC-AE first encoded the motion and appearance of selected objects and then clustered the training samples into normal clusters. An object is considered anomalous in the inference stage if the one-versus-rest classifier's highest classification score is negative. Its extension, the Background-Agnostic framework [40], introduced instance segmentation, allowing the model to focus only on the primary object. In addition, the authors used pseudo-anomaly examples to perform adversarial learning to improve the appearance and motion auto-encoders.

To make full use of the contextual information, Yu *et al.* [176] proposed a Video Event Completion (VEC) method that used appearance and motion as cues to locate regions of interest. VEC recovered the original video events by solving visual completion tests to capture high-level semantics and inferring deleted patches. Georgescu *et al.* [39] designed several self-supervised learning tasks, including discrimination of forward/backward moving objects, discrimination of objects in continuous/intermittent frames, and reconstruction of object-specific appearance. In the testing phase, anomalous objects would lead to large prediction discrepancies.

Doshi *et al.* [33] proposed an Online Anomaly Detection (OAD) scheme that used detected object information such as location, category, and size as input to a clustering model to detect anomalous events. HF^2-VAD [92] seamlessly integrated frames reconstruction and prediction. It used memory to record the normal pattern of optical flow reconstruction and captured the correlation between RGB frames and optical flow using a conditional variation auto-encoder.

Chen *et al.* [15] proposed a Bidirectional Prediction (BiP) architecture with three consistency constraints. Specifically, prediction consistency considered the symmetry of motion and appearance in forward and backward prediction. Association consistency considered the correlation between frames and optical flow, and temporal consistency was used to ensure that BiP can generate temporally consistent frames. The Hierarchical Scene Normality-Binding Modeling (HSNBM) framework in [5] attempted to parse global and local surveillance scenes. They proposed a scene object-binding frame prediction module to model the relationship between foreground and background.

4 WEAKLY-SUPERVISED ABNORMAL EVENT DETECTION

Using weakly semantic video-level labels to supervise the model was first proposed by Sultani *et al.* [144] in 2018, laying the foundation for WAED based on multiple instance learning. The synchronously released UCF-crime dataset collected 13 classes of real-world criminal behaviors and provided video-level labels for training sets. Following researchers [149, 194] made UVAD datasets meet WAED requirements by moving some anomalous test videos to the training set, introducing various WAED benchmarks such as the reorganized UCSD Ped2 [88] and ShanghaiTech Weakly [73, 194]. In 2020, the XD-Violence [168] dataset extended GVAED from unimodal video analysis to multimodal signal processing. Since WAED no longer follows the basic AD assumptions, it has long been underappreciated. WAED model only works when anomalies can be predefined, and enough positive samples can be collected. However, the development of WAED datasets [144, 168, 194] and methods [35, 88, 149], and their excellent performance in traffic accident detection [102] and

violence detection [163, 169] suggest that WAED has become a mainstream GVAED technical route. The anomalies in specific application scenarios, and the anomalous behaviors of concern to humans, are not unbounded, so it is feasible to predefine and collect positive samples for training WAED models. In contrast to UVAD which treats all unseen events as anomalies, WAED focuses only on predefined, human-self-aware anomalies. Therefore, its anomalous response results are more informative in complex environments [89]. This section will elaborate on the existing WAED models and classify the existing methods into unimodal [35, 73, 91, 144, 149, 183, 194] and multimodal [163, 168, 169, 178] models according to the input data modalities, which helps to guide the GVAED development from video processing to multimodal understanding community.

4.1 Unimodal Models

The unimodal models are similar to UVAD in terms of input data, with the difference that the former computes the anomaly score directly, while the latter needs first to learn normality and detect anomalies in a downstream task. Specifically, the unimodal WAED model typically slices the unedited video into several fixed-size clips. They consider each clip as an instance, and all clips from a video form a bag with the same video-level label. And then, pre-trained feature extractors, such as Convolutional 3D (C3D) [150], Temporal Segment Networks (TSN) [160], and Inflated 3D (I3D) [11], is used to extract the spatial-temporal features of the examples. Generally, the scoring module takes deep representations as input and calculates the anomaly score for each instance with the supervision of video-level labels. We present the feature extractors, decision logic, and main contributions of existing unimodal methods in Table 6.

The MIL ranking framework [144] introduced multiple instance learning to GVAED for the first time, using a 3-layer Fully Connected Network (FCN) to predict high anomaly scores for anomalous clips and introducing sparsity and smooth constraints to avoid drastic fluctuations in the score curve. Zhu and Newsam [198] considered motion as the key to WAED performance. To this end, they proposed a temporal augmented network to learn motion-aware features and use attention blocks to incorporate temporal context into a MIL ranking model. Snehashis *et al.* [106] used a dual-stream CNN to extract spatial and temporal features separately and fed the fused features as spatial-temporal representations into an FCN to perform anomaly classification. The authors compared the performance of different deep CNN architectures (e.g., ResNet- 50 [49], Inception V3 [146], and VGG-16 [143]) for feature extraction.

Zhong *et al.* [194] treated WAED as a supervised learning task under noisy labels, arguing that the supervised action recognition models can perform anomaly detection after the label noise is removed. In response, they designed a graph convolutional network to correct the labels. [157] proposed Anomaly Regression Network (ARNet) to learn discriminative features WAED. Specifically, ARNet used dynamic multiple-instance learning loss and center loss to enlarge the inter-class distance instances and reduce the intra-class distance of regular instances, respectively.

Waseem *et al.* [74] proposed a two-stage WAED framework that first used an echo state network to obtain spatially and temporally aware features. And then, they used a 3D convolutional network to extract spatial-temporal features and fuse them with the features from the first stage as the input to a binary classifier. Tian *et al.* proposed Robust Temporal Feature Magnitude (RTFM) learning by training a feature volume learning function to identify positive examples efficiently. In addition, RTFM utilized self-attention to capture both long and short-time correlations. Muhammad *et al.* [183] proposed a self-reasoning framework that uses binary clustering to generate pseudo-labels to supervise the MIL regression models.

The CLustering Assisted Weakly Supervised (CLAWS) learning with normalcy sppression in [181] proposed a random batch-based training strategy to reduce the correlation between batches. In addition, the authors introduced a loss based on clustering distance to optimize the network to weaken the effect of label noise. Ammar *et al.* [60] proposed a Deep Temporal Coding-Decoding (DTED) to capture the temporal evolution of videos over time. They treated instances

Table 6. Unimodal WAED Models.

Year	Method	Fearure	Decision Logic	Contributions
2018	MIR [144]	C3D	MIL ranking	Proposing to use video-level labels to supervise a MIL regression model to compute frame-level anomaly scores.
2019	TS-CNN [106]	ResNet-50, VGG-16	Binary classification	Using dual-stream CNNs to extract spatial and temporal features from video frames and optical flow.
2019	GCLNC [194]	C3D, TSN^{RGB}, TSN$^{Optical\ flow}$	Action classification	Treating WAED as a supervised task under noise labels and using graph convolutional networks to correct noise labels.
2019	MAF [198]	VGG16, C3D, Inception, I3D	MIL ranking	Proposing a temporal-enhanced network to learn motion-aware features for MIL ranking model.
2020	ARNet [157]	I3D^{RGB}, I3D$^{Optical\ flow}$, I3D^{conc}	MIL ranking	Designing dynamic multi-instance learning loss and center loss for expanding the inter-class distance and reducing the intra-class distance of normal instances.
2020	SRF [183]	C3D	MIL ranking	Using a clustering algorithm to generate binary pseudo-labels to assist the training of regression networks.
2020	CLAWS [181]	C3D	MIL ranking	Proposing a random batch-based training strategy to reduce the correlation between batches.
2021	RTFM [149]	C3D^{RGB}, I3D^{RGB}	MIL ranking	Proposing RTFM to explore the important temporal correlations for efficient identification of positive instances.
2021	WSAL [102]	I3D	MIL ranking	Fusing spatial and temporal contexts to perform weakly-supervised video anomaly localization while proposing an enhancement strategy to eliminate noise interference.
2021	MIST [35]	C3D^{RGB}, I3D^{RGB}	MIL ranking	Using a pseudo-label generator to generate reliable frame-level labels and extract task-specific deep representations.
2022	DTED [60]	C3D	MIL ranking	Proposing a deep temporal coding scheme to capture the temporal evolution of video examples over time, reducing the false alarm rate of anomaly detection.
2022	WSTR [186]	I3D^{RGB}	MIL ranking	Exploring the temporal relationships between video clips, and capturing the task-specific features.
2022	GCLNC+ [73]	C3D, TSN^{RGB}, TSN$^{Optical\ flow}$	MIL ranking	Presenting a graph convolutional network for cleaning label noise with integrated feature similarity and temporal consistency of anomaly analysis.
2022	CNN-echo	CNN	Binary classification	Using 2D convolutional networks and echo state networks to obtain local ratio representations, and then using 3D convolutional networks to extract spatial-temporal features.
2022	SMR [88]	I3D	MIL ranking	Using RNNs to capture temporal correlations and using a clustering algorithm to generate pseudo-labels to aid the training of the MIL regression module.
2022	STA [91]	C3D^{RGB}, I3D^{RGB}	MIL ranking	Proposing recurrent criss-cross attention to explore the connection between local spatial-temporal representations.

of the same bag as sequential visual data rather than as independent individuals. In addition, DTED uses joint loss to optimize to maximize the average distance between normal and abnormal videos.

The Weakly-supervised Temporal Relationship (WSTR) learning framework [186] enhanced the model's discriminative power by exploring the temporal relationships between clips. The proposed transformer-enabled encoder converts

Table 7. Multimodal WAED Models.

Year	Method	Input Modality	Contributions
2020	HL-Net [168]	Video + Audio	Collecting the XD-Violence violence detection datasets and proposing a three-branch neural network model for multimodal anomaly detection.
2021	FVAI [117]	Video + Audio	Proposing a pooling-based feature fusion strategy to fuse video and audio information to obtain more discriminative feature representations.
2022	SC [119]	Video + Audio	Proposing an audio-visual scene classification dataset containing 5 classes of anomalous events and a deep classification model.
2022	MACIL-SD [178]	Video + Audio	Proposing a modality-aware contrastive instance learning with a self-distillation strategy to address the modality heterogeneity challenges.
2022	ACF [164]	Video + Audio	Proposing a two-stage multimodal information fusion method for violence detection that first refines video-level labels into clip-level labels.
2022	MSAF [163]	Video + Audio, Video + Optical flow	Proposing multimodal labels refinement to refine video-level ground truth into pseudo-clip-level labels and implicitly align multimodal information with multimodal supervise-attention fusion network.
2022	MD [138]	Video + Audio + Flow	Using mutual distillation to transfer information and proposing a multi-modal fusion network to fuse video, audio, and flow features.
2022	HL-Net+ [169]	Video + Audio	Introducing coarse-grained violent frame and fine-grained violent event detection tasks and proposing a network for audio-visual violence detection.
2022	AGAN [120]	Video + Audio	Using cross-modal interaction to enhance video and audio and computing high-confidence violence scores using temporal convolution.

the task-irrelevant representations into task-specific features by mining the semantic correlations and positional relationships between video clips. Weakly Supervised Anomaly Localization (WSAL) [102] performed anomaly detection by fusing temporal and spatial contexts and proposed a higher-order context encoding model to measure temporal dynamic changes. In addition, the authors collected a dataset called TAD for traffic anomaly detection.

Feng *et al.* [35] proposed a Multi-Instance Self-Training (MIST) framework consisting of a multi-instance pseudo label generator and a self-guided attention-enhancing feature encoder for generating more reliable fragment-level pseudo labels and extracting task-specific representations, respectively. Liu *et al.* [88] proposed a Self-guiding Multi-instance Ranking (SMR) framework that used a clustering module to generate pseudo labels to aid the training of supervised multi-instance regression models to explore task-relevant feature representations. The authors compared the performance of different recurrent neural networks in exploring temporal correlation. Spatial-Temporal Attention (STA) [91] explored the connection between example local representations and global spatial-temporal features through a recurrent cross-attention operation and used mutual cosine loss to encourage the enhanced features to be task specific.

4.2 Multimodal Models

Multimodal GVAED [163, 168, 169] aims to mine effective clues related to anomalies from various data, such as video, audio, and optical flow. Real-world data are heterogeneous, and effectively exploiting the complementary nature of multimodal data is the key to developing robust and efficient GVAED models. Due to the limitation of datasets, most existing works [168, 169, 178] focused on video and audio information fusion to detect violent behaviors from surveillance videos. Moreover, inspired by the frame-level multi-stream UVAD models [76, 87], recent work [163]

considered RGB frames and optical flow as different modalities. We display the modalities and principles of existing multimodal GVAED models [117, 119, 120, 138, 163, 164, 168, 169, 178] in Table 7.

Wu *et al.* [168] released the first multimodal GVAED dataset and proposed a three-branch network called HL-Net for multimodal violence detection. Specifically, the similarity branch used a similarity prior to capture long-range correlations. In contrast, the proximity branch used proximity prior to capture local location relationships, and the scoring branch dynamically captured the proximity of predicted scores. Experimental results demonstrated the multimodal data's positive impact on GVAED. The following MACIL-SD in [178] utilized a lightweight dual-stream network to overcome the heterogeneity challenge. It used self-distillation to transfer unimodal visual knowledge to audio-visual models to narrow the semantic gap between multimodal features.

Researchers [117, 120] attempted to explore more effective feature extraction and multimodal information fusion strategies. For example, Pang *et al.* [117] proposed to use a bilinear pooling mechanism to fuse visual and audio information and encourage the model to learn from each other to obtain a more effective representation. Audio-Guided Attention Network (AGAN) [120] first used a deep neural network to extract video and audio features and then enhanced the features in the temporal dimension using a cross-modal perceptual local arousal network.

Wei *et al.* [164] proposed a two-stage multimodal information fusion method, which first refines video-level hard labels into clip-level soft labels and then uses an attention module for multimodal information fusion. Their extension work, Multimodal Supervised Attentional Augmentation Fusion (MSAF) [163], used attention fusion to align information and achieved implicit alignment of multimodal data.

Shang *et al.* [138] observed that existing models are limited by small datasets and proposed mutual distillation to transfer information from large-scale datasets to small datasets. They proposed a multimodal attention fusion strategy to fuse RGB images, audio, and flow to obtain a more discriminative representation. [119] introduced an audio-visual scene classification task and released a multimodal dataset. The authors try different deep networks and fusion strategies to explore the most effective classification model.

5 FULLY-UNSUPERVISED VIDEO ANOMALY DETECTION

Fully-unsupervised Video Anomaly Detection (FVAD) does not limit the composition of the training data and requires no data annotation. In other words, FVAD tries to learn an anomaly detector from the random raw data, which is a newly emerged technical route in recent years.

Ionescu *et al.* [152] introduced the unmasking technique to computer vision tasks, proposing an FVAD framework that requires no training sequences. They iteratively trained a binary classifier to distinguish two consecutive video sequences and simultaneously removed the most discriminative features at each step. Inspired by [152], Liu *et al.* [85] tried to establish the connection between heuristic unmasking and multiple classifiers two sample tests to improve its testing capability. In this regard, they proposed a history sampling method to increase the testing power as well as to improve the GVAED performance. Li *et al.* [77] first used a distribution clustering framework to identify the possible anomalous samples in the training data, and then used the clustered subset of normal data to train the auto-encoder. An encoder that can describe normality was obtained by repeating normal subset selection and representation learning.

The recent representative FVAD works are Self-trained Deep Ordinal regression (SDOR) [116] and Generative Cooperative Learning (GCL) [182], which attempted to learn anomaly scorers from unlabeled videos in an end-to-end manner, as shown in Fig. 8(a) and 8(b). Specifically, SDOR [116] first determined the initial pseudo-normal and abnormal sets and then computed the abnormal scores using pre-trained ResNet-50 and FCN. The representation module and the scorer were optimized iteratively in a self-training manner. Moreover, Lin *et al.* [82] looked at the pseudo label

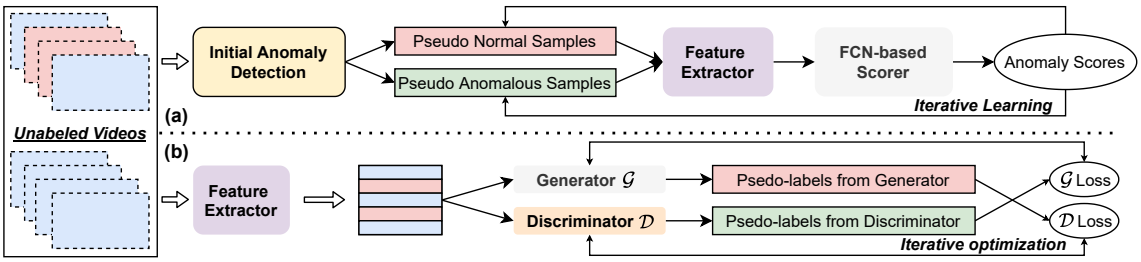

Fig. 8. Workflow of two representative FVAD methods: (a) SDOR [116] and (b) GCL [182]. Given the unlabeled videos, the SDOR first divided them into pseudo-normal and anomalous sets by initial anomaly detection. GCL introduces cross-supervision to train the generator $\mathcal{G}$ and discriminator $\mathcal{D}$ to learn anomaly detectors. The pseudo-labels from $\mathcal{G}$ and $\mathcal{D}$ are used to compute each other's losses.

generation process in SDOR from a causal inference perspective and proposed a causal graph to analyze confounding effects and eliminate the impact of noisy pseudo labels. In addition, their proposed CIL model improved significantly by performing counterfactual inference to capture long-range temporal dependencies.

In contrast, GCL [182] attempted to exploit the low-frequency nature of anomalous events. It included a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$, which were supervised by each other in a cooperative rather. The generator primarily generated representations for normal events. While for anomaly events, the generator used negative learning techniques to distort the anomaly representation and generated pseudo-labels to train $\mathcal{D}$. The discriminator estimated the probability of anomalies and created pseudo labels to improve $\mathcal{G}$. Hu *et al.* [51] also exploited the rarity of anomalies, with the idea that The small number of anomalous samples in the training set has a limited impact on the normality of the model learning. Inspired by the Masked Auto-Encoder (MAE) [47], their proposed TMAE learned representations using a visual transformer performing a complementary task. Since MAE [47] only applied masks in 2D images, while video anomalies are related to the temporal information, TMAE first located video foregrounds and constructed temporal cubes to be masked objects.

6 SUPERVISED VIDEO ANOMALY DETECTION

Supervised video anomaly detection requires frame-level or pixel-level labels to supervise models to distinguish between normal and anomalies. Therefore, it is often considered a classification task rather than a mainstream GVAED scheme. On the one hand, collecting fine-grained labeled anomalous samples is time-consuming. On the other hand, the anomalous behavior occurs gradually, and the degree of anomaly is a relative value, while manual labeling can only provide discrete 0/1 labels, which cannot adequately describe the severity and temporal continuity of video anomalies. Existing SVAD methods usually consider VAED a binary classification task under data imbalance conditions. However, game engines can simulate various types of anomalous events and provide frame-level and pixel-level personalized annotations. With the penetration of synthetic datasets in vision tasks, supervised training of GVAED models with virtual anomalies is expected to become possible. Researchers need to focus on the domain adaptation problem posed by synthetic datasets, i.e., how to cope with the covariate shifts between synthetic data and the real-world surveillance video and the ensuing performance degradation. Moreover, although the training set contains partially labeled anomalies, SVAD models still need to consider how to reasonably generalize the anomalies to detect unseen anomalous events in real-world scenarios. SVAD is an open-set recognition task rather than a supervised binary classification.

Table 8. EER and AUC Comparison of Early Unsupervised Methods on Benchmark Datasets.

Year	Method	Ped1 AUC	Ped1 EER	Ped2 AUC	Ped2 EER	Avenue AUC	Avenue EER	ShanghaiTech AUC
2015	DRAM [170]	92.1	16.0	90.8	17.0	-	-	-
2015	STVP [3]	93.9	12.9	94.6	10.6	-	-	-
2016	CMAC [189]	85.0	-	90.0	-	-	-	-
2016	FF-AE [46]	81.0	27.9	90.0	21.7	70.2	25.1	60.9
2017	DEM [37]	92.5	15.1	-	-	-	-	-
2017	CFS [68]	82.0	21.1	84.0	19.2	-	-	-
2017	WTA-AE [151]	91.9	15.9	92.8	11.2	82.1	24.2	-
2017	EBM [156]	70.3	35.4	86.4	16.5	78.8	27.2	-
2017	CPE [153]	78.2	24.0	80.7	19.0	-	-	-
2017	LDGK [50]	-	-	92.2	13.9	-	-	-
2017	sRNN [98]	-	-	92.2	-	81.7	-	68.0
2017	GANS [126]	97.4	8.0	93.5	14.0	-	-	-
2017	OGNG [145]	93.8	-	94.0	-	-	-	-
2018	FFP [84]	83.1	-	95.4	-	85.1	-	72.8
2018	PP-CNN [125]	95.7	8.0	88.4	18.0	-	-	-
2019	FAED [95]	93.8	14.0	95.0	-	-	-	-
2019	NNC [58]	-	-	-	-	88.9	-	-
2019	OC-AE [57]	-	-	97.8	-	90.4	-	84.9
2019	AMC [114]	-	-	96.2	-	86.9	-	-
2019	MLR [155]	82.3	23.5	99.2	2.5	71.5	36.4	-
2019	memAE [42]	-	-	94.1	-	83.3	-	71.2
2019	MLEP [83]	-	-	-	-	92.8	-	76.8
2019	BMAN [66]	-	-	96.6	-	90.0	-	76.2
2020	Street Scene [122]	77.3	25.9	88.3	18.9	72.0	33.0	-
2020	IPR [147]	82.6	-	96.2	-	83.7	-	73.0
2020	DFSN [123]	86.0	23.3	94.0	14.1	87.2	18.8	-

Table 9. AUC Comparison of Recent Unsupervised and *Fully-unsupervised (Marked in Italics)* Methods on Benchmark Datasets.

Year	Method	Ped2	Avenue	ShanghaiTech	Year	Method	Ped2	Avenue	ShanghaiTech
2020	MNAD-R [118]	90.2	82.8	69.8	2020	MNAD-P [118]	97.0	88.5	70.5
2020	DD-GAN [31]	95.6	84.9	73.7	2020	*SDOR* [116]	83.2	-	-
2020	ASSAD [32]	97.8	86.4	71.6	2020	FSSA [96]	96.2	85.8	77.9
2020	VEC [176]	97.3	89.6	74.8	2020	Multispace[56]	95.4	86.8	73.6
2020	CDD-AE [14]	96.5	86.0	73.3	2021	CDD-AE+ [13]	96.7	87.1	73.7
2021	Multi-task (object level) [39]	99.8	91.9	89.3	2021	Multi-task (frame level) [39]	92.4	86.9	83.5
2021	Multi-task (late fusion) [39]	99.8	92.8	90.2	2021	HF^{2}AVD [92]	99.3	91.1	76.2
2021	AST-AE [86]	96.6	85.2	68.8	2021	ROADMAP[162]	96.3	88.3	76.6
2021	CT-D2GAN[36]	97.2	85.9	77.7	2022	AMAE [87]	97.4	88.2	73.6
2022	STM-AE [90]	98.1	89.8	73.8	2022	BiP [15]	97.4	86.7	73.6
2022	AR-AE [64]	98.3	90.3	78.1	2022	TAC-Net[56]	98.1	88.8	77.2
2022	STC-Net [192]	96.7	87.8	73.1	2022	HSNBM [5]	95.2	91.6	76.5
2022	*CIL(ResNet50)+DCFD* [82]	97.9	85.9	-	2022	*CIL(ResNet50)+DCFD+CTCE* [82]	99.4	87.3	-
2022	*CIL(I3D-RGB)+DCFD+CTCE* [82]	98.7	90.3	-	2022	*GCL$_{PT}$ (RESNEXT)* [182]	-	-	78.93

7 PERFORMANCE COMPARISON

We collect the performance of existing works on publicly available datasets [78, 84, 94, 144, 194] to quantitatively compare the superiority and present the GVAED development progress. Table 8 presents the frame-level AUC and EER of the early UVAD models on UCSD Ped1 & Ped2 [78], and CUHK Avenue [94] datasets and the frame-level AUC on the ShanghaiTech [84] dataset. Since the recent UVAD [14, 42, 87, 90] and FVAD [82, 116, 182] only report frame-level AUC as the main evaluation metric, we have collated these methods separately in Table 9. The ShanghaiTech dataset was proposed in 2018 with the FFP [84] model, so methods before this time were usually tested without this dataset. With the advantage of its data size and quality, ShanghaiTech has become the most widely used UVAD benchmark. An interesting phenomenon is that the object-level methods outperform other frame-level and patch-level models on the cross-scene ShanghaiTech dataset. For example, the frame-level AUC of the Multi-task [39] model is as high as 90.2%, which is 12.1% higher than the state-of-the-art frame-level methods [64]. It shows that for cross-scene GVAED tasks, using an object detector to separate the foreground object of interest from the scene can effectively avoid interference

Table 10. Quantitative Performance Comparison of Weakly-supervised Methods on Public Datasets.

Method	Feature	UCF-Crime AUC	UCF-Crime FAR	ShanghaiTech AUC	ShanghaiTech FAR
MIR [144]	C3D^{RGB}	75.40	1.90	86.30	0.15
TCN [187]	C3D^{RGB}	78.70	-	82.50	0.10
Zhong [194]	C3D^{RGB}	80.67	3.30	76.44	-
ARNet [157]	C3D^{RGB}	-	-	85.01	0.57
	I3D^{RGB}	-	-	85.38	0.27
	I3D$^{RGB+Optical\ Flow}$	-	-	91.24	0.10
MIST [35]	C3D^{RGB}	81.40	2.19	93.13	1.71
	I3D^{RGB}	82.30	0.13	94.83	0.05
RTFM [149]	C3D^{RGB}	83.28	-	91.51	-
	I3D^{RGB}	84.30	-	97.21	-
SMR [88]	I3D$^{RGB+Optical\ Flow}$	81.70	-	-	-
DTED [60]	C3D^{RGB}	79.49	0.50	87.42	-

of the background. In addition, the multi-stream model learns normality in both temporal and spatial dimensions and generally outperforms the single-stream model. The usage frequency shows that UCSD Ped2 [78], CUHK Avenue [94], and ShanghaiTech [84] have become the prevailing benchmarks for UVAD evaluation. Future work should consider testing and comparing the proposed methods on these three datasets.

Table 10 presents the performance of WAED methods on the UCF-Cirme [144] and ShanghaiTech weaky [194] datasets. As mentioned previously, WAED models usually rely on pre-trained feature extractors [11, 150, 160] to obtain feature representations. Commonly used features include C3D^{RGB}, I3D^{RGB}, and I3D$^{RGB+Optical\ flow}$. The performance gap of the same model using different features show that the effectiveness of the WAED model is related to the pre-trained feature extractors, with the I3D outperforming the simple $3 \times 3 \times 3$ convolution-based C3D network due to the separate consideration of temporal information variation. Future WAED work should test the performance of the proposed model on current commonly used features or provide the performance of existing works on emerging features to demonstrate that the performance gain comes from the model design rather than benefiting from a more robust feature extraction network. In addition to detection performance, other metrics are processing speed and deployment cost. Unfortunately, due to differences in implementation platforms, we cannot quantitatively compare the parameter size and inference speed of the existing works. We recommend that future work provide related metrics on mainstream computing platforms.

8 CHALLENGES AND TRENDS

8.1 Research Challenges

8.1.1 Mock Anomalies vs. Real Anomalies. GVAED aims to automatically detect anomalous events in the living environment to provide a safe space for humans. However, the difficulty of collecting anomalies makes most of the existing datasets formulate abnormal events by human simulation, such as the UMN [25], CUHK Avenue [94], and ShanghaiTech [84]. The mock anomalies are simpler, and their spatial-temporal patterns differ significantly from normal events, resulting in well-trained models difficult to detect complex anomalies. In addition, the set of limited categories of anomalous events conflicts with the diverse nature of real anomalies. As a result, models learned on such datasets

perform poorly in real-world scenarios. Therefore, collecting datasets containing various real anomalies and designing models to bridge the gap between mock and real anomalies is an essential challenge for GVAED development.

8.1.2 Single-scene vs. Multi-scenes. Mainstream unsupervised datasets [78, 94] and UVAD methods [14, 84, 118] only consider single-scene videos, while the real world always contains multiple scenes, which constitutes another challenge for UVAD methods. Although the UMN [25] and ShanghaiTech [84] datasets include multiple scenes, the anomalous events of the former are all crowd dispersal, while the 13 scenes of the latter are similar. Therefore, most UVAD methods [84, 90] do not consider the scene differences but learn normality directly from the original video as in other single-scene datasets [78, 94]. Recent researchers [40, 57] believe that object-level methods are a feasible way to learn scene-invariant normality by extracting specific objects from the scenes and then analyzing the spatial-temporal patterns of objects and backgrounds separately. The multi-scene problem is inescapable for model deployment as it is almost impossible to train a scene-specific model for each terminal device. Developing cross-scene GVAED models using domain adaptation/generalization techniques [165] to learn scene-invariant normality is a definite challenge.

8.1.3 Real data vs. Synthetic data. Due to the rarity and diversity of anomalies, collecting and labeling anomalous events is time-consuming and laborious. Therefore, researchers [1] have considered using game engines [34, 137] to synthesize anomaly data. We remain optimistic about this attempt and believe it may lead to new research opportunities for GVAED. While anomaly detection tasks suffer from a lack of data and missing labels, synthetic data can generate various anomalous samples and provide precise frame-level or even pixel-level annotations, making it possible to develop SVAD models and save data preparation costs for large-scale GVAED model training. However, a concomitant challenge is that covariate shifts between synthetic and real data may make the trained GVAED models not work in real scenes.

8.1.4 Unimodal vs. Multimodal. Researchers [168, 169] are aware of the positive impact of multimodal data (e.g., audio) for GVAED. However, existing works are stuck on the lack of datasets and the validity of model structures. XD-Violence [168] is the only mainstream multimodal GVAED dataset, but it only contains video and audio, and much data is collected from movies and games rather than the real world. With the popularity of IoT, using various sensors to collect environmental information (e.g., temperature, brightness, and humidity) can assist cameras in detecting abnormal events. However, mining useful clues from valid data and developing multimodal efficient GVAED models need further research.

8.1.5 Single-view vs. Multi-view. In places such as traffic intersections and parks, the same area is usually covered by multiple camera views, deriving another task: anomalous event detection in multi-view videos. Multi-view data can provide more comprehensive environmental awareness data, which is wildly used for re-identification [81, 175], tracking [148] and gaze estimation [80]. However, existing datasets [25, 78, 84, 94] are all single-view, making multi-view GVAED research still a gap. The simplest idea is to combine data from all views to train the same model and determine anomalies through a voting or winner-take-all strategy. However, such approaches are training-costly and do not consider the differences and complementarities between multi-view data. Therefore, multi-view GVAED remains to be investigated.

8.1.6 Off-line vs. On-line Detection. The deployable Intelligent video surveillance systems (IVSS) need to process continuously generated video streams 24/7 on-line and respond to anomalous events in real-time so that noteworthy clips can be saved in time to reduce storage and transmission costs. Unfortunately, existing GVAED models are designed for public datasets rather than real-time video streams, primarily pursuing detection performance while avoiding the on-line detection challenges. For example, the dominant prediction-based methods [84, 87, 90, 118] in UVAD route can only give the prediction error of the current input in a single-step execution, while the Informative anomaly score needs are

obtained after performing the maximum-minimum normalization over the prediction error of all frames. Although the model can directly determine the current frame as an anomaly with a pre-set error threshold, existing attempts show that the manually selected threshold is unreliable. WAED [35, 88, 102, 144, 149] can directly output anomaly scores for segments. However, the input to the scoring module is a discriminative spatial-temporal representation rather than the original video. The representations usually rely on pre-trained 3D convolution-based feature extractors [11, 150, 160]. The time cost is unacceptable for resource-limited terminal devices. Therefore, developing on-line detection models is the primary challenge for GVAED deployment, determining its application potential in IVSS and streaming media platforms.

8.2 Development Trends

8.2.1 Data level. From single-scene [78, 94] to multi-scene [84, 122], from real-word vides [25] to synthetic data [1], and from unimodal [144] to multimodal [168], GVAED datasets are moving towards large-scale and realistic scenarios. We see this as a positive trend that will continue with the growth of on-line video platforms and tools. On the one hand, real-world scenarios and anomalies are diverse, so efficient models for real-world applications need to be trained on large-scale datasets that contain various anomalous events. On the other hand, the Internet has made it possible to collect multi-scene and multi-view videos, including sufficient rare anomalous behaviors such as violence and crime. Furthermore, multimodal and synthetic data will be increasingly important in GVAED research. The XD-Violence [168] dataset has demonstrated the positive impact of multimodal data on GVAED. In the future, with streaming media (e.g., TikTok, Netflix, and Hulu) and on-line video sites (e.g., YouTube, iQIYI, and Youku), more modal data can be collected. Besides, virtual game engines (e.g., Airsim [137] and Carla [34]) can synthesize rare anomalous events and provide fine-grained annotations on demand.The connection of GVAED with other tasks (e.g., multimodal analysis [163] and few-shot learning [96]) will tend to be close, with the latter inspiring the design of GVAED models under specific data conditions.

8.2.2 Representation level. Deep learning has enabled spatial-temporal representations to be derived directly from the raw videos in an end-to-end manner without a human prior. The earlier deep GVAED models benefit from CNNs and pursue complex deep networks to extract more abstract features. For example, the UVAD models attempt to introduce dual-stream networks to learn spatial and temporal representations [14, 87, 90], and use 3D convolutional networks to model temporal features [42, 193]. From C3D [150] to I3D [11], the WAED models [35, 88, 149] benefit from more powerful pre-trained feature extractors and achieves general performance gains on existing datasets [144, 194]. We observe that the representation means of WAED will become increasingly sophisticated. New visual representation learning models such as Transformer [4, 54, 75, 154] will drive WAED development. In contrast, UVAD does not pursue abstract representations. Overly powerful deep networks may lead to missing anomalous events as normal due to overgeneralization [42, 118]. Future researchers should consider using clever representation strategies to balance the model's powerful representation of normal events and the limited generalization of abnormal events. Powerful generative models such as graph learning [55, 99, 107] and diffusion models [24] are expected to provide more effective normality learning tools for UVAD. In addition, researchers should consider introducing emerging techniques (e.g., domain adaptation [40, 159]) to develop GVAED models for learning scene-invariant representation from multi-scene and multi-view videos.

8.2.3 Deployment level. Model deployment is an inevitable trend for GVAED development. As mentioned above, the multi-scene and the diversity of anomalies in real-world videos pose new challenges for model design and training, such as on-line detection, lightweight models, and high view robustness. On the one hand, the computational resources of terminal devices are limited. Most deep GVAED methods are overly pursuing performance at the expense of running costs. On the other hand, existing models are trained off-line, which cannot perform real-time detection. Model compression

[29] and knowledge distillation [43] can drive the development of lightweight GVAED models. On-line evolutive learning [70, 71] will enable models to optimize learnable parameters in complex working environments dynamically.

8.2.4 Methodology level. This review compares the four main GVAED technical routes: UVAD, WAED, SVAD, and FVAD. UVAD has been regarded as the mainstream solution, although WAED gradually dominates in recent years. However, the trend of UVAD is unclear due to its performance saturation on limited datasets [78]. In addition, the setting of anomalies in UVAD datasets makes UVAD models challenging to work in complex scenes. Self-supervised visual representation technicals (e.g., contrast learning [18, 44, 48, 56] and deep clustering [10, 185]) may provide new ideas for UVAD. In contrast, WAED has been widely noticed as a research hotspot due to its excellent performance in crime detection [144]. In addition, the multimodal video anomaly detection tasks also follow WAED routes. SVAD is once abandoned due to the lack of labels and anomalies. However, it may face new research opportunities with the emergence of synthetic datasets [1]. In contrast, FVAD can learn directly from raw video data without the cost of training data filtering and annotations, making it a hot research topic. The various routes are not completely independent, and existing works [89, 166] have started to combine the assumptions of different methods to develop more efficient GVAED models.

9 CONCLUSION

This review is the first to integrate the deep learning-driven technical routes based on different assumptions and learning frameworks into a unified generalized video anomaly event detection framework. We provide a hierarchical GVAED taxonomy that systematically organizes the existing literature by supervision, input data, and network structure, focusing on the recent advances such as weakly-supervised, fully-unsupervised, and multimodal methods. To provide a comprehensive review of the extant work, we collect benchmark datasets and available codes, sort out the development lines of various methods, and perform performance comparisons and strengths analysis. This survey helps clarify the connections among deep GVAED routes and advance community development. In addition, we analyze research challenges and future trends in the context of deep learning technology development and possible problems faced by GAED model deployment, which can serve as a guide for future researchers and engineers.

ACKNOWLEDGMENTS

This work is funded by the China Mobile Research Fund of the Chinese Ministry of Education (Grant No. KEH2310029). This work is also supported in part by the Shanghai Key Research Lab. of NSAI and in part by the Joint Lab. on Networked AI Edge Computing Fudan University-Changan.

REFERENCES

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. 2022. UBnormal: New Benchmark for Supervised Open-Set Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20143–20153.

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence* 30, 3 (2008), 555–560.

[3] Borislav Antić and BjörnT28 Ommer. 2015. Spatio-temporal Video Parsing for Abnormality Detection. *arXiv preprint arXiv:1502.06235* (2015).

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.

[5] Qianyue Bao, Fang Liu, Yang Liu, Licheng Jiao, Xu Liu, and Lingling Li. 2022. Hierarchical scene normality-binding modeling for anomaly detection in surveillance videos. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6103–6112.

[6] Yannick Benezeth, P-M Jodoin, Venkatesh Saligrama, and Christophe Rosenberger. 2009. Abnormal events detection based on spatio-temporal co-occurences. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2458–2465.

[7] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. 2021. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–33.

[8] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, and Zhifeng Hao. 2021. Appearance-motion memory consistency network for video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 938–946.

[9] Yiheng Cai, Jiaqi Liu, Yajun Guo, Shaobin Hu, and Shinan Lang. 2021. Video anomaly detection with multi-scale feature and temporal information fusion. *Neurocomputing* 423 (2021), 264–273.

[10] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems* 33 (2020), 9912–9924.

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

[12] S Chandrakala, K Deepak, and G Revathy. 2022. Anomaly detection in surveillance videos: a thematic taxonomy of deep models, review and performance analysis. *Artificial Intelligence Review* (2022), 1–50.

[13] Yunpeng Chang, Zhigang Tu, Wei Xie, Bin Luo, Shifu Zhang, Haigang Sui, and Junsong Yuan. 2021. Video anomaly detection with spatio-temporal dissociation. *Pattern Recognition* 122 (2021), 108213.

[14] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. 2020. Clustering driven deep autoencoder for video anomaly detection. In *European Conference on Computer Vision*. Springer, 329–345.

[15] Chengwei Chen, Yuan Xie, Shaohui Lin, Angela Yao, Guannan Jiang, Wei Zhang, Yanyun Qu, Ruizhi Qiao, Bo Ren, and Lizhuang Ma. 2022. Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection. In *Proceedings of the American association for artificial intelligence*. 1–9.

[16] Dongyue Chen, Pengtao Wang, Lingyi Yue, Yuxin Zhang, and Tong Jia. 2020. Anomaly detection in surveillance video based on bidirectional prediction. *Image and Vision Computing* 98 (2020), 103915.

[17] Dongyue Chen, Lingyi Yue, Xingya Chang, Ming Xu, and Tong Jia. 2021. NM-GAN: Noise-modulated generative adversarial network for video anomaly detection. *Pattern Recognition* 116 (2021), 107969.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[19] Zhaoyu Chen, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Wenqiang Zhang. 2022. Towards Practical Certifiable Patch Defense with Vision Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15148–15158.

[20] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. 2015. Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2909–2917.

[21] Yong Shean Chong and Yong Haur Tay. 2017. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*. Springer, 189–196.

[22] Peter Christiansen, Lars N Nielsen, Kim A Steen, Rasmus N Jørgensen, and Henrik Karstoft. 2016. DeepAnomaly: Combining background subtraction and deep learning for detecting obstacles and anomalies in an agricultural field. *Sensors* 16, 11 (2016), 1904.

[23] Andrew A Cook, Göksel Mısırlı, and Zhong Fan. 2019. Anomaly detection for IoT time-series data: A survey. *IEEE Internet of Things Journal* 7, 7 (2019), 6481–6494.

[24] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2022. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747* (2022).

[25] Xinyi Cui, Qingshan Liu, Mingchen Gao, and Dimitris N Metaxas. 2011. Abnormal detection using interaction energy potentials. In *CVPR 2011*. IEEE, 3161–3167.

[26] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.

[27] Navneet Dalal, Bill Triggs, and Cordelia Schmid. 2006. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*. Springer, 428–441.

[28] K Deepak, S Chandrakala, and C Krishna Mohan. 2021. Residual spatiotemporal autoencoder for unsupervised video anomaly detection. *Signal, Image and Video Processing* 15, 1 (2021), 215–222.

[29] Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proc. IEEE* 108, 4 (2020), 485–532.

[30] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. 2005. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance*. IEEE, 65–72.

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

[32] Keval Doshi and Yasin Yilmaz. 2020. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 934–935.

[33] Keval Doshi and Yasin Yilmaz. 2021. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition* 114 (2021), 107865.

[34] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*. PMLR, 1–16.

[35] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng. 2021. Mist: Multiple instance self-training framework for video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14009–14018.

[36] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. 2021. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5546–5554.

[37] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu. 2017. Learning deep event models for crowd anomaly detection. *Neurocomputing* 219 (2017), 548–556.

[38] Félix Fuentes-Hurtado, Abdolrahim Kadkhodamohammadi, Evangello Flouty, Santiago Barbarisi, Imanol Luengo, and Danail Stoyanov. 2019. EasyLabels: weak labels for scene segmentation in laparoscopic videos. *International journal of computer assisted radiology and surgery* 14, 7 (2019), 1247–1257.

[39] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2021. Anomaly detection in video via self-supervised and multi-task learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12742–12752.

[40] Mariana Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. 2021. A background-agnostic framework with adversarial training for abnormal event detection in video. *IEEE transactions on pattern analysis and machine intelligence* 44, 9 (2021), 4505–4523.

[41] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

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

[43] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

[44] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[45] H Haberfehlner, AI Buizer, KL Stolk, SS van de Ven, I Aleo, LA Bonouvrié, J Harlaar, and MM van der Krogt. 2020. Automatic video tracking using deep learning in dyskinetic cerebral palsy. *Gait Posture* 81 (2020), 132–133.

[46] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. 2016. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 733–742.

[47] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16000–16009.

[48] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[50] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE international conference on computer vision*. 3619–3627.

[51] Jingtao Hu, Guang Yu, Siqi Wang, En Zhu, Zhiping Cai, and Xinzhong Zhu. 2022. Detecting Anomalous Events from Unlabeled Videos via Temporal Masked Auto-Encoding. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[52] Xing Hu, Shiqiang Hu, Yingping Huang, Huanlong Zhang, and Hanbing Wu. 2016. Video anomaly detection using deep incremental slow feature analysis network. *IET Computer Vision* 10, 4 (2016), 258–267.

[53] Xing Hu, Yingping Huang, Xiumin Gao, Lingkun Luo, and Qianqian Duan. 2018. Squirrel-cage local binary pattern and its application in video anomaly detection. *IEEE Transactions on Information Forensics and Security* 14, 4 (2018), 1007–1022.

[54] Chao Huang, Chengliang Liu, Jie Wen, Lian Wu, Yong Xu, Qiuping Jiang, and Yaowei Wang. 2022. Weakly Supervised Video Anomaly Detection via Self-Guided Temporal Discriminative Transformer. *IEEE Transactions on Cybernetics* (2022).

[55] Chao Huang, Yabo Liu, Zheng Zhang, Chengliang Liu, Jie Wen, Yong Xu, and Yaowei Wang. 2022. Hierarchical Graph Embedded Pose Regularity Learning via Spatio-Temporal Transformer for Abnormal Behavior Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 307–315.

[56] Chao Huang, Zhihao Wu, Jie Wen, Yong Xu, Qiuping Jiang, and Yaowei Wang. 2021. Abnormal event detection using deep contrastive learning for intelligent video surveillance system. *IEEE Transactions on Industrial Informatics* 18, 8 (2021), 5171–5179.

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

[58] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. 2019. Detecting abnormal events in video using narrowed normality clusters. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1951–1960.

[59] Sabah Abdulazeez Jebur, Khalid A Hussein, Haider Kadhim Hoomod, Laith Alzubaidi, and José Santamaría. 2022. Review on Deep Learning Approaches for Anomaly Event Detection in Video Surveillance. *Electronics* 12, 1 (2022), 29.

[60] Ammar Mansoor Kamoona, Amirali Khodadadian Gostar, Alireza Bab-Hadiashar, and Reza Hoseinnezhad. 2023. Multiple instance-based video anomaly detection using deep temporal encoding–decoding. *Expert Systems with Applications* 214 (2023), 119079.

[61] B Ravi Kiran, Dilip Mathew Thomas, and Ranjith Parakkal. 2018. An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging* 4, 2 (2018), 36.

[62] Kwang-Eun Ko and Kwee-Bo Sim. 2018. Deep convolutional framework for abnormal behavior detection in a smart surveillance system. *Engineering Applications of Artificial Intelligence* 67 (2018), 226–234.

[63] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. 2008. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

[64] Viet-Tuan Le and Yong-Guk Kim. 2023. Attention-based residual autoencoder for video anomaly detection. *Applied Intelligence* 53, 3 (2023), 3240–3254.

[65] Jooyeon Lee, Woo-Jeoung Nam, and Seong-Whan Lee. 2022. Multi-Contextual Predictions with Vision Transformer for Video Anomaly Detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 1012–1018.

[66] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. 2019. BMAN: bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing* 29 (2019), 2395–2408.

[67] Roberto Leyva, Victor Sanchez, and Chang-Tsun Li. 2017. The LV dataset: A realistic surveillance video dataset for abnormal event detection. In *2017 5th international workshop on biometrics and forensics (IWBF)*. IEEE, 1–6.

[68] Roberto Leyva, Victor Sanchez, and Chang-Tsun Li. 2017. Video anomaly detection with compact feature sets for online performance. *IEEE Transactions on Image Processing* 26, 7 (2017), 3463–3478.

[69] Daoheng Li, Xiushan Nie, Xiaofeng Li, Yu Zhang, and Yilong Yin. 2022. Context-related video anomaly detection via generative adversarial network. *Pattern Recognition Letters* 156 (2022), 183–189.

[70] Di Li and Liang Song. 2022. Multi-Agent Multi-View Collaborative Perception Based on Semi-Supervised Online Evolutive Learning. *Sensors* 22, 18 (2022), 6893.

[71] Di Li, Xiaoguang Zhu, and Liang Song. 2022. Mutual match for semi-supervised online evolutive learning. *Applied Intelligence* (2022), 1–15.

[72] Nanjun Li, Faliang Chang, and Chunsheng Liu. 2020. Spatial-temporal cascade autoencoder for video anomaly detection in crowded scenes. *IEEE Transactions on Multimedia* 23 (2020), 203–215.

[73] Nannan Li, Jia-Xing Zhong, Xiujun Shu, and Huiwen Guo. 2022. Weakly-supervised anomaly detection in video surveillance via graph convolutional label noise cleaning. *Neurocomputing* 481 (2022), 154–167.

[74] Nannan Li, Jia-Xing Zhong, Xiujun Shu, and Huiwen Guo. 2022. Weakly-supervised anomaly detection in video surveillance via graph convolutional label noise cleaning. *Neurocomputing* 481 (2022), 154–167.

[75] Shuo Li, Fang Liu, and Licheng Jiao. 2022. Self-training multi-sequence learning with Transformer for weakly supervised video anomaly detection. *Proceedings of the AAAI, Virtual* 24 (2022).

[76] Tong Li, Xinyue Chen, Fushun Zhu, Zhengyu Zhang, and Hua Yan. 2021. Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection. *Neurocomputing* 439 (2021), 256–270.

[77] Tangqing Li, Zheng Wang, Siying Liu, and Wen-Yan Lin. 2021. Deep unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3636–3645.

[78] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. 2013. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence* 36, 1 (2013), 18–32.

[79] Yuanyuan Li, Yiheng Cai, Jiaqi Liu, Shinan Lang, and Xinfeng Zhang. 2019. Spatio-Temporal Unity Networking for Video Anomaly Detection. *IEEE Access* 7 (2019), 172425–172432. https://doi.org/10.1109/ACCESS.2019.2954540

[80] Dongze Lian, Lina Hu, Weixin Luo, Yanyu Xu, Lixin Duan, Jingyi Yu, and Shenghua Gao. 2018. Multiview multitask gaze estimation with deep convolutional neural networks. *IEEE transactions on neural networks and learning systems* 30, 10 (2018), 3010–3023.

[81] Weipeng Lin, Yidong Li, Xiaoliang Yang, Peixi Peng, and Junliang Xing. 2019. Multi-view learning for vehicle re-identification. In *2019 IEEE international conference on multimedia and expo (ICME)*. IEEE, 832–837.

[82] Xiangru Lin, Yuyang Chen, Guanbin Li, and Yizhou Yu. 2022. A Causal Inference Look at Unsupervised Video Anomaly Detection. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 1620–1629. https://ojs.aaai.org/index.php/AAAI/article/view/20053

[83] Wen Liu, Weixin Luo, Zhengxin Li, Peilin Zhao, Shenghua Gao, et al. 2019. Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies.. In *IJCAI*. 3023–3030.

[84] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. 2018. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6536–6545.

[85] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. 2018. Classifier Two Sample Test for Video Anomaly Detections.. In *BMVC*. 71.

[86] Yang Liu, Shuang Li, Jing Liu, Hao Yang, Mengyang Zhao, Xinhua Zeng, Wei Ni, and Liang Song. 2021. Learning Attention Augmented Spatial-temporal Normality for Video Anomaly Detection. In *2021 3rd International Symposium on Smart and Healthy Cities (ISHC)*. IEEE, 137–144.

[87] Yang Liu, Jing Liu, Jieyu Lin, Mengyang Zhao, and Liang Song. 2022. Appearance-Motion United Auto-Encoder Framework for Video Anomaly Detection. *IEEE Transactions on Circuits and Systems II: Express Briefs* 69, 5 (2022), 2498–2502.

[88] Yang Liu, Jing Liu, Wei Ni, and Liang Song. 2022. Abnormal Event Detection with Self-guiding Multi-instance Ranking Framework. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 01–07.

[89] Yang Liu, Jing Liu, Mengyang Zhao, Shuang Li, and Liang Song. 2022. Collaborative Normality Learning Framework for Weakly Supervised Video Anomaly Detection. *IEEE Transactions on Circuits and Systems II: Express Briefs* 69, 5 (2022), 2508–2512.

[90] Yang Liu, Jing Liu, Mengyang Zhao, Dingkang Yang, Xiaoguang Zhu, and Liang Song. 2022. Learning Appearance-Motion Normality for Video Anomaly Detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[91] Yang Liu, Jing Liu, Xiaoguang Zhu, Donglai Wei, Xiaohong Huang, and Liang Song. 2022. Learning Task-Specific Representation for Video Anomaly Detection with Spatial-Temporal Attention. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2190–2194.

[92] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13588–13597.

[93] Vina Lomte, Satish Singh, Siddharth Patil, Siddheshwar Patil, and Durgesh Pahurkar. 2019. A Survey on Real World Anomaly Detection in Live Video Surveillance Techniques. *International Journal of Research in Engineering, Science and Management* 2, 2 (2019), 2581–5792.

[94] Cewu Lu, Jianping Shi, and Jiaya Jia. 2013. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*. 2720–2727.

[95] Cewu Lu, Jianping Shi, Weiming Wang, and Jiaya Jia. 2019. Fast abnormal event detection. *International Journal of Computer Vision* 127, 8 (2019), 993–1011.

[96] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. 2020. Few-shot scene-adaptive anomaly detection. In *European Conference on Computer Vision*. Springer, 125–141.

[97] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 439–444.

[98] Weixin Luo, Wen Liu, and Shenghua Gao. 2017. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE international conference on computer vision*. 341–349.

[99] Weixin Luo, Wen Liu, and Shenghua Gao. 2021. Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection. *Neurocomputing* 444 (2021), 332–337.

[100] Weixin Luo, Wen Liu, Dongze Lian, and Shenghua Gao. 2021. Future frame prediction network for video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[101] Weixin Luo, Wen Liu, Dongze Lian, Jinhui Tang, Lixin Duan, Xi Peng, and Shenghua Gao. 2019. Video anomaly detection with sparse coding inspired deep neural networks. *IEEE transactions on pattern analysis and machine intelligence* 43, 3 (2019), 1070–1084.

[102] Hui Lv, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. 2021. Localizing anomalies from weakly-labeled videos. *IEEE transactions on image processing* 30 (2021), 4505–4515.

[103] Ke Ma, Michael Doescher, and Christopher Bodden. 2015. Anomaly detection in crowded scenes using dense trajectories. *University of Wisconsin-Madison* (2015).

[104] Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z Sheng, Hui Xiong, and Leman Akoglu. 2021. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[105] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. 2010. Anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1975–1981.

[106] Snehashis Majhi, Ratnakar Dash, and Pankaj Kumar Sa. 2020. Two-Stream CNN architecture for anomalous event detection in real world scenarios. In *International Conference on Computer Vision and Image Processing*. Springer, 343–353.

[107] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. 2020. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10539–10547.

[108] Jefferson Ryan Medel and Andreas Savakis. 2016. Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv preprint arXiv:1612.00390* (2016).

[109] Harshadkumar S Modi, Dr Parikh, and A Dhaval. 2022. A Survey on Crowd Anomaly Detection. *International Journal of Computing and Digital Systems* 12, 1 (2022), 1081–1096.

[110] Ruwan Nawarathna, JungHwan Oh, Jayantha Muthukudage, Wallapak Tavanapong, Johnny Wong, Piet C De Groen, and Shou Jiang Tang. 2014. Abnormal image detection in endoscopy videos using a filter bank and local binary patterns. *Neurocomputing* 144 (2014), 70–91.

[111] Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, and Xinghuo Yu. 2019. Spatiotemporal anomaly detection using deep learning for real-time video surveillance. *IEEE Transactions on Industrial Informatics* 16, 1 (2019), 393–402.

[112] Rashmiranjan Nayak, Umesh Chandra Pati, and Santos Kumar Das. 2021. A comprehensive review on deep learning-based methods for video anomaly detection. *Image and Vision Computing* 106 (2021), 104078.

[113] Khac-Tuan Nguyen, Dat-Thanh Dinh, Minh N Do, and Minh-Triet Tran. 2020. Anomaly detection in traffic surveillance videos with gan-based future frame prediction. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 457–463.

[114] Trong-Nguyen Nguyen and Jean Meunier. 2019. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1273–1283.

[115] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2021. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.

[116] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. 2020. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12173–12182.

[117] Wen-Feng Pang, Qian-Hua He, Yong-jian Hu, and Yan-Xiong Li. 2021. Violence detection in videos based on fusing visual and audio information. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2260–2264.

[118] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. 2020. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14372–14381.

[119] Lam Pham, Dat Ngo, Tho Nguyen, Phu Nguyen, Truong Hoang, and Alexander Schindler. 2022. An audio-visual dataset and deep learning frameworks for crowded scene classification. In *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*. 23–28.

[120] Yujiang Pu and Xiaoyu Wu. 2022. Audio-Guided Attention Network for Weakly Supervised Violence Detection. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE, 219–223.

[121] Rohit Raja, Prakash Chandra Sharma, Md Rashid Mahmood, and Dinesh Kumar Saini. 2022. Analysis of anomaly detection in surveillance video: recent trends and future vision. *Multimedia Tools and Applications* (2022), 1–17.

[122] Bharathkumar Ramachandra and Michael Jones. 2020. Street Scene: A new dataset and evaluation protocol for video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2569–2578.

[123] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. 2020. Learning a distance function with a Siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2598–2607.

[124] Bharathkumar Ramachandra, Michael J Jones, and Ranga Raju Vatsavai. 2020. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence* 44, 5 (2020), 2293–2312.

[125] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. 2018. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1689–1698.

[126] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. 2017. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE international conference on image processing (ICIP)*. IEEE, 1577–1581.

[127] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. 2019. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1896–1904.

[128] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[129] Khosro Rezaee, Sara Mohammad Rezakhani, Mohammad R Khosravi, and Mohammad Kazem Moghimi. 2021. A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. *Personal and Ubiquitous Computing* (2021), 1–17.

[130] Mehrsan Javan Roshtkhari and Martin D Levine. 2013. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer vision and image understanding* 117, 10 (2013), 1436–1452.

[131] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. 2015. Real-time anomaly detection and localization in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 56–62.

[132] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. 2017. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing* 26, 4 (2017), 1992–2004.

[133] Venkatesh Saligrama and Zhu Chen. 2012. Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on computer vision and pattern recognition*. IEEE, 2112–2119.

[134] Venkatesh Saligrama, Janusz Konrad, and Pierre-Marc Jodoin. 2010. Video anomaly identification. *IEEE Signal Processing Magazine* 27, 5 (2010), 18–33.

[135] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, and Partha Pratim Roy. 2020. Anomaly detection in road traffic using visual surveillance: A survey. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–26.

[136] Sam Sattarzadeh, Mahesh Sudhakar, and Konstantinos N Plataniotis. 2021. SVEA: A Small-scale Benchmark for Validating the Usability of Post-hoc Explainable AI Solutions in Image and Signal Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4158–4167.

[137] Shital Shah, Debadeepta Dey, Chris Lovett, and Ashish Kapoor. 2018. Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics*. Springer, 621–635.

[138] Yimeng Shang, Xiaoyu Wu, and Rui Liu. 2022. Multimodal Violent Video Recognition Based on Mutual Distillation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 623–637.

[139] Md Sharif, Lei Jiao, Christian W Omlin, et al. 2022. Deep Crowd Anomaly Detection: State-of-the-Art, Challenges, and Future Research Directions. *arXiv preprint arXiv:2210.13927* (2022).

[140] Prakhar Singh and Vinod Pankajakshan. 2018. A Deep Learning Based Technique for Anomaly Detection in Surveillance Videos. In *2018 Twenty Fourth National Conference on Communications (NCC)*. 1–6. https://doi.org/10.1109/NCC.2018.8599969

[141] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. 2017. Deep appearance features for abnormal behavior detection in video. In *International Conference on Image Analysis and Processing*. Springer, 779–789.

[142] Liang Song, Xing Hu, Guanhua Zhang, Petros Spachos, Konstantinos N Plataniotis, and Hequan Wu. 2022. Networking systems of ai: on the convergence of computing and communications. *IEEE Internet of Things Journal* 9, 20 (2022), 20352–20381.

[143] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. *Advances in neural information processing systems* 28 (2015).

[144] Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6479–6488.

[145] Qianru Sun, Hong Liu, and Tatsuya Harada. 2017. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition* 64 (2017), 187–201.

[146] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[147] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. 2020. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters* 129 (2020), 123–130.

[148] Zheng Tang, Renshu Gu, and Jenq-Neng Hwang. 2018. Joint multi-view people tracking and pose estimation for 3D scene reconstruction. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[149] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. 2021. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4975–4986.

[150] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[151] Hanh TM Tran and David Hogg. 2017. Anomaly detection using a convolutional winner-take-all autoencoder. In *Proceedings of the British Machine Vision Conference 2017*. British Machine Vision Association.

[152] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. 2017. Unmasking the abnormal events in video. In *Proceedings of the IEEE international conference on computer vision*. 2895–2903.

[153] Francesco Turchini, Lorenzo Seidenari, and Alberto Del Bimbo. 2017. Convex polytope ensembles for spatio-temporal anomaly detection. In *International Conference on Image Analysis and Processing*. Springer, 174–184.

[154] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

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

[156] Hung Vu, Dinh Phung, Tu Dinh Nguyen, Anthony Trevors, and Svetha Venkatesh. 2017. Energy-based models for video anomaly detection. *arXiv preprint arXiv:1708.05211* (2017).

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

[158] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision* 103, 1 (2013), 60–79.

[159] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[160] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[161] Tian Wang, Meina Qiao, Zhiwei Lin, Ce Li, Hichem Snoussi, Zhe Liu, and Chang Choi. 2018. Generative neural networks for anomaly detection in crowded scenes. *IEEE Transactions on Information Forensics and Security* 14, 5 (2018), 1390–1399.

[162] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. 2021. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE transactions on neural networks and learning systems* (2021).

[163] Donglai Wei, Yang Liu, Xiaoguang Zhu, Jing Liu, and Xinhua Zeng. 2022. MSAF: Multimodal Supervise-Attention Enhanced Fusion for Video Anomaly Detection. *IEEE Signal Processing Letters* 29 (2022), 2178–2182.

[164] Dong-Lai Wei, Chen-Geng Liu, Yang Liu, Jing Liu, Xiao-Guang Zhu, and Xin-Hua Zeng. 2022. Look, Listen and Pay More Attention: Fusing Multi-Modal Information for Video Violence Detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1980–1984.

[165] Garrett Wilson and Diane J Cook. 2020. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 5 (2020), 1–46.

[166] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. 2022. Self-supervised Sparse Representation for Video Anomaly Detection. In *European Conference on Computer Vision*. Springer, 729–745.

[167] Peng Wu, Jing Liu, and Fang Shen. 2019. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE transactions on neural networks and learning systems* 31, 7 (2019), 2609–2622.

[168] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*. Springer, 322–339.

[169] Peng Wu, Xiaotao Liu, and Jing Liu. 2022. Weakly supervised audio-visual violence detection. *IEEE Transactions on Multimedia* (2022).

[170] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553* (2015).

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

[172] Hang Xu, Lewei Yao, Wei Zhang, Xiaodan Liang, and Zhenguo Li. 2019. Auto-fpn: Automatic network architecture adaptation for object detection beyond classification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6649–6658.

[173] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.

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

[175] Qingze Yin, Guodong Ding, Shaogang Gong, Zhenmin Tang, et al. 2021. Multi-view label prediction for unsupervised learning person re-identification. *IEEE Signal Processing Letters* 28 (2021), 1390–1394.

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

[177] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. 2021. Abnormal event detection and localization via adversarial event prediction. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[178] Jiashuo Yu, Jinyu Liu, Ying Cheng, Rui Feng, and Yuejie Zhang. 2022. Modality-Aware Contrastive Instance Learning with Self-Distillation for Weakly-Supervised Audio-Visual Violence Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*. 6278–6287.

[179] Hongchun Yuan, Zhenyu Cai, Hui Zhou, Yue Wang, and Xiangzhi Chen. 2021. TransAnomaly: Video Anomaly Detection Using Video Vision Transformer. *IEEE Access* 9 (2021), 123977–123986.

[180] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. 2020. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14183–14193.

[181] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. 2020. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *European Conference on Computer Vision*. Springer, 358–376.

[182] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. 2022. Generative Cooperative Learning for Unsupervised Video Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14744–14754.

[183] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. 2020. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters* 27 (2020), 1705–1709.

[184] Cheng Zhan, Han Hu, Zhi Wang, Rongfei Fan, and Dusit Niyato. 2019. Unmanned aircraft system aided adaptive video streaming: A joint optimization approach. *IEEE Transactions on Multimedia* 22, 3 (2019), 795–807.

[185] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. 2020. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6688–6697.

[186] Dasheng Zhang, Chao Huang, Chengliang Liu, and Yong Xu. 2022. Weakly Supervised Video Anomaly Detection via Transformer-Enabled Temporal Relation Learning. *IEEE Signal Processing Letters* (2022).

[187] Jiangong Zhang, Laiyun Qing, and Jun Miao. 2019. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 4030–4034.

[188] Qianqian Zhang, Guorui Feng, and Hanzhou Wu. 2022. Surveillance video anomaly detection via non-local U-Net frame prediction. *Multimedia Tools and Applications* (2022), 1–16.

[189] Ying Zhang, Huchuan Lu, Lihe Zhang, and Xiang Ruan. 2016. Combining motion and appearance cues for anomaly detection. *Pattern Recognition* 51 (2016), 443–452.

[190] Zhenzhen Zhang, Jianjun Hou, Qinglong Ma, and Zhaohong Li. 2015. Efficient video frame insertion and deletion detection based on inconsistency of correlations between local binary pattern coded frames. *Security and Communication networks* 8, 2 (2015), 311–320.

[191] Zhe Zhang, Shiyao Ma, Zhaohui Yang, Zehui Xiong, Jiawen Kang, Yi Wu, Kejia Zhang, and Dusit Niyato. 2022. Robust semi-supervised federated learning for images automatic recognition in internet of drones. *IEEE Internet of Things Journal* (2022).

[192] Mengyang Zhao, Yang Liu, Jing Liu, and Xinhua Zeng. 2022. Exploiting Spatial-temporal Correlations for Video Anomaly Detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 1727–1733.

[193] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. 2017. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*. 1933–1941.

[194] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1237–1246.

[195] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. 2019. Anomalynet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security* 14, 10 (2019), 2537–2550.

[196] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[197] Shifu Zhou, Wei Shen, Dan Zeng, Mei Fang, Yuanwang Wei, and Zhijiang Zhang. 2016. Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes. *Signal Processing: Image Communication* 47 (2016), 358–368.

[198] Yi Zhu and Shawn Newsam. 2019. Motion-aware feature for improved video anomaly detection. *arXiv preprint arXiv:1907.10211* (2019).