Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 11, No. 2, 2024

A Survey on Self-Supervised Learning-Based Video Anomaly Detection

Mengjie Hu, Qingtao Wu

School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China

Abstract: Video anomaly detection (VAD) exhibits promising applications across diverse domains, bolstering intelligence, security, and operational efficiency, thereby catalyzing industry growth. This paper begins by examining the research background and significance of VAD, providing an in-depth analysis of its relevance across various sectors. Subsequently, from a machine learning standpoint, recent advancements in self-supervised learning (SSL)-based VAD models are systematically categorized and summarized, elucidating their underlying principles and deployment scenarios. Additionally, commonly utilized datasets in VAD are introduced to facilitate readers' understanding of model assessment and comparative analysis. Lastly, discussions on future trajectories and extant challenges in VAD are undertaken to foster deeper exploration and propel the advancement of this domain.

Keywords: Video Anomaly Detection (VAD); Self-Supervised Learning; Machine Learning Standpoint; Commonly Utilized Datasets in VAD; Promote the Future Development of VAD.

1. Introduction

In recent years, anomaly detection technology has garnered significant attention in the field of computer vision. As it continues to evolve, there is a pressing need for this technology across various domains to drive technological innovation. In the education sector, it can be utilized to monitor student behavior and identify and intervene in bullying incidents. In the retail industry, analyzing customer behavior can optimize store layout and enhance the shopping experience. The financial services sector can leverage anomaly detection technology to monitor and prevent fraudulent activities, thereby improving transaction security. Multiple industries are actively exploring the integration of this technology into their operations and security systems to enhance efficiency, ensure safety, and even develop new business models.

VAD [1] aims to identify abnormalities in videos, such as unusual object behaviors, atypical motion patterns, abnormal visual features, or anomalous events. In surveillance systems, analyzing video streams captured by surveillance cameras enables the timely detection of anomalies such as intrusion, theft, and violence [2], thereby enhancing security in both public and private domains. In the industrial sector, VAD can monitor anomalies on production lines, including equipment malfunctions and production line halts [3], facilitating timely intervention to improve production efficiency and reliability. In traffic monitoring systems, it can detect anomalies such as traffic accidents, congestion, and traffic violations [4], thereby enhancing the efficiency and safety of road traffic management. Overall, with the development of computer vision and machine learning technologies, research and applications of VAD are becoming increasingly important. However, the challenges and issues facing VAD are illustrated in Fig 1.

This paper primarily summarizes existing research achievements in VAD based on self-supervised learning. It provides a detailed introduction from the perspectives of VAD methods and benchmark datasets and discusses prospects and directions for future research.

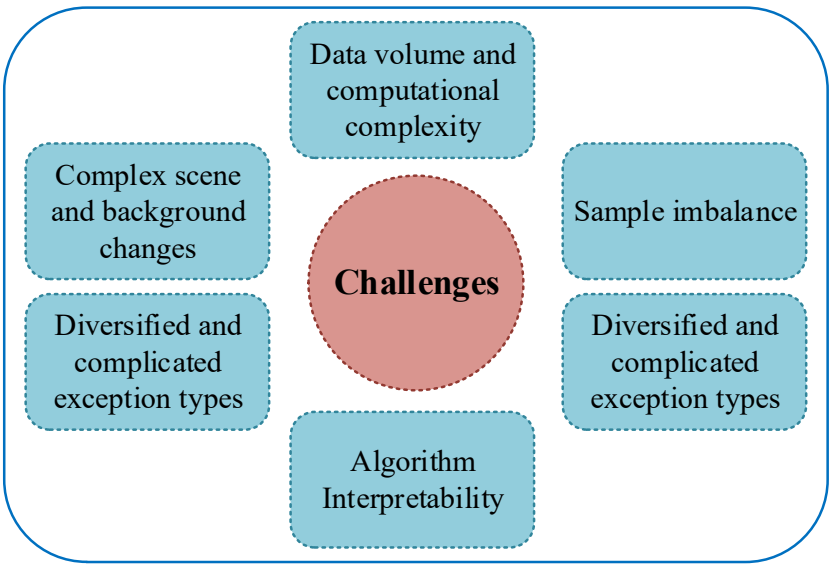

Fig 1. Problems and challenges faced by VAD

2. VAD based on Self-supervised Learning

Self-supervised learning aims to more effectively utilize data for model training, reducing the human and time costs of tasks such as VAD. The methods include: (1) Reconstruction error. (2) Spatio-temporal consistency. (3) Self-supervised contrastive learning. (4) Generative models.

2.1. Reconstruction Error

The reconstruction error method utilizes autoencoders or Generative Adversarial Network (GAN) to map video sequences to low-dimensional representations and then map them back to the original video sequences. Anomaly detection is based on the error between the reconstructed video and the original video, with anomalies typically resulting in larger errors.

Recently, Wang [5] proposed the Memory-Augmented Appearance and Motion Network (MAAM-Net) to address the challenge of constrained defect boundaries. MAAM-Net utilizes frame reconstruction and flow prediction, reassembling normal events through memory modules to handle unknown events and introducing margin-based latent losses. Xing [6] introduced a histogram error estimation module to eliminate adverse errors, thus improving anomaly localization performance without increasing costs. Hyun [7] introduced alienation and assimilation strategies to diversify normal spatial and temporal information by learning fundamental patterns of multi-level latent representations, thereby enhancing detection and localization performance.

While the reconstruction error method performs well in some cases, it also has limitations. For example, it is susceptible to background changes, which may lead to false positives or false negatives when the background of the video sequence changes. If the anomaly pattern is similar to the normal pattern, the reconstruction error may not reflect the presence of anomalies. Additionally, the reconstruction error method typically requires a large amount of training data to learn representations. Therefore, when choosing a method, it is necessary to consider the specific application scenarios and data characteristics.

2.2. Spatiotemporal Consistency

The spatiotemporal consistency method utilizes the temporal and spatial relationships between frames in video sequences for anomaly detection. By predicting the next frame, the spatio-temporal consistency model is learned, and then this model is utilized to detect frames that do not conform to consistency, which are considered anomalies.

Hao [8] proposed a Spatio-Temporal Consistency Enhancement Network (STCEN) aimed at highlighting the interference of anomalous data by addressing both spatial and temporal aspects, employing a resampling module to widen the score gap between normal and anomalous content. Li [9] introduced a Memory-Augmented Spatio-Temporal Consistency Network, aiming to model the potential consistency between spatial appearance and temporal motion by learning a unified spatio-temporal representation. They introduced a spatio-temporal memory fusion module to record spatio-temporal prototypes of regular patterns from the unified spatio-temporal representation, thus increasing the gap between normal and anomalous events in the feature space.

While the spatio-temporal consistency method has advantages in VAD, it also has some limitations. For example, it is sensitive to background changes, difficult to handle complex anomaly patterns, has high computational complexity, difficulties in selecting model parameters, and is limited by local features.

2.3. Self-supervised Contrastive Learning

The self-supervised contrastive learning method involves comparing frames or segments within a video sequence with themselves or other frames or segments for learning. Normal frames or segments should be similar, while abnormal ones are less similar. Thus, contrastive learning methods can be employed for anomaly detection.

Chen [10] proposed the Contrastive Action Representation Learning (CARL) framework, which includes a frame-by-frame video encoder considering spatio-temporal relationships and introduces Sequence Contrastive Loss (SCL). This method optimizes the embedding space by minimizing the Kullback-Leibler (KL) divergence between the sequence similarity of two enhanced views and the prior Gaussian distribution of timestamp distance, as shown in Fig 2. Dwibedi [11] proposed a self-supervised representation learning method based on temporal alignment tasks between videos, utilizing Temporal Cycle Consistency (TCC) to train the network. This method can search for temporal correspondences across multiple videos and generate embeddings for each frame, simply matching frames to align videos using the nearest neighbors in the learned embedding space. Dvornik [12] addressed the problem of sequence-to-sequence alignment containing outliers, proposing the Drop-DTW algorithm, which aligns common signals between sequences while automatically removing outlier elements from the matching. Its effectiveness as a training loss in various applications has been validated through experiments. Wang [13] introduced an object-centric feature alignment method, injecting local object features into verb and noun branches and proposing a symbiotic attention mechanism to encourage mutual interaction between the two branches, selecting candidates most relevant to the action for classification.

From the aforementioned studies, it is evident that contrastive learning requires a large number of paired labels, including pairs of similar and dissimilar samples. However, obtaining a large-scale dataset for contrastive learning poses a significant challenge for training limited contrastive models. Additionally, different sequences may have varying lengths, and such disparities in sequence lengths could pose challenges for model learning. Furthermore, computational efficiency is a noteworthy concern when dealing with large-scale datasets.

2.4. Generative Model

The generative model evaluates the anomaly level of an input video sequence by learning the probability distribution of video sequences. Anomalies cause deviations between the input video sequence and the distribution of normal data learned by the generative model.

Huang [14] proposed the Self-Supervised Attention Generative Adversarial Network (SSAGAN), comprising a self-attention predictor, a regular discriminator, and a self-supervised discriminator. However, SSAGAN reduces the model's generalization ability to anomalous frames, leading

to increased detection errors for such frames. Chen [15]introduced an end-to-end pipeline (NM-GAN), combining an encoder-decoder reconstruction network and a CNN-based discriminator within a GAN-like framework. This method trains the discriminator network to differentiate between abnormal and normal samples using reconstruction error maps. Wang [16] presented a shallow generative network to obtain a model resembling a Gaussian mixture to fit the distribution of actual data. This approach demonstrates improved performance in detecting both local and global anomaly events.

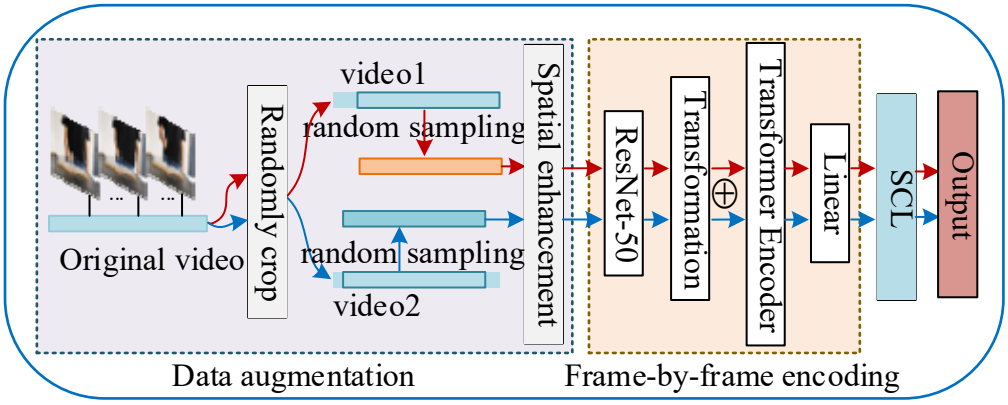

Fig 2. Comparative action representation learning (CARL) framework

The drawbacks of generative model methods include high computational complexity, difficulty in model training, and sample imbalance. Therefore, it is necessary to combine other methods to improve detection performance.

3. Benchmark Datasets

3.1. UCSD Dataset

The UCSD dataset [17] records various daily activities and abnormal events, such as crowd gatherings and traffic congestion. This dataset is divided into two subsets, each corresponding to different scenarios. The Peds1 subset includes clips of crowds walking towards and away from the camera, along with some perspective distortion, comprising 34 training video samples and 36 testing video samples. The other subset, Peds2, contains scenes of pedestrians moving along the camera plane, with a total of 16 training video samples and 12 testing video samples. Each recorded video segment in each scenario contains approximately 200 frames.

3.2. UCF-Crime Dataset

The UCF-Crime dataset [18] is a large-scale dataset consisting of 128 hours of video footage, comprising 1,900 continuous untrimmed surveillance videos. This dataset encompasses 13 types of real-life abnormal behaviors, including abuse, arrest, arson, assault, road accidents, burglary, explosion, fighting, robbery, shooting, theft, shoplifting, and vandalism.

3.3. ShanghaiTech Campus Dataset

The ShanghaiTech dataset [19] collected video sequences from the campus of ShanghaiTech University, capturing various daily activities and abnormal events such as traffic accidents and crowd gatherings. This dataset comprises 330 normal videos for training and 107 abnormal videos for testing, encompassing 13 types of abnormal events.

3.4. Avenue Dataset

The Avenue dataset [20] covers video sequences from various scenarios, including campus and street environments. Some of the video sequences in this dataset contain various abnormal events such as crowd gatherings and unusual movements. The dataset comprises 16 training videos and 21 testing videos, totaling 30,652 frames (15,328 frames for training and 15,324 frames for testing).

3.5. Street Scene Dataset

The Street Scene dataset [21] consists of 46 training segments and 35 testing segments, with a resolution of 1280×720. These segments were captured in a dual-carriageway setting containing both bike lanes and sidewalks.

4. Prospect

VAD, as an important research field, holds vast prospects for future development. This paper delves into the current state of research in VAD based on self-supervised learning, focusing on both anomaly detection methods and benchmark datasets. Overall, significant progress has been made in the field of VAD. Future research can explore more diverse and comprehensive self-supervised learning tasks to better leverage the intrinsic information in video data, as well as consider integrating cross-modal learning tasks. Furthermore, VAD technology can be more widely applied in various domains such as smart cities, intelligent transportation, and industrial production, providing tailored solutions for different application scenarios.

Acknowledgments

This work was supported in part by the Key Technologies R & D Program of Henan Province under Grant No. 222102210080 and 242102211024, in part by the Longmen Laboratory Frontier Exploration Project of Henan Province under Grant No. MQYTSKT035, in part by the joint Funds for Science and Technology Research and Development Plan of Henan Province under Grant No. 222103810031.

References

[1] Xiang T, Gong S. Video behavior profiling for anomaly detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2008, 30(5): 893-908.

[2] Sánchez F L, Hupont I, Tabik S, et al. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly

detection, crowd emotions, datasets, opportunities and prospects[J]. Information Fusion, 2020, 64: 318-335.

[3] Xia X, Pan X, Li N, et al. GAN-based anomaly detection: A review[J]. Neurocomputing, 2022, 493: 497-535.

[4] Santhosh K K, Dogra D P, Roy P P. Anomaly detection in road traffic using visual surveillance: A survey[J]. ACM Computing Surveys (CSUR), 2020, 53(6): 1-26.

[5] Wang L, Tian J, Zhou S, et al. Memory-augmented appearance-motion network for video anomaly detection[J]. Pattern Recognition, 2023, 138: 109335.

[6] Xing P, Li Z. Visual anomaly detection via partition memory bank module and error estimation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[7] Hyun W, Nam W J, Lee S W. Dissimilate-and-assimilate strategy for video anomaly detection and localization[J]. Neurocomputing, 2023, 522: 203-213.

[8] Hao Y, Li J, Wang N, et al. Spatiotemporal consistency-enhanced network for video anomaly detection[J]. Pattern Recognition, 2022, 121: 108232.

[9] Li Z, Zhao M, Zeng X, et al. Memory-Augmented Spatial-Temporal Consistency Network for Video Anomaly Detection [C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Singapore: Springer Nature Singapore, 2023: 95-107.

[10] Chen M, Wei F, Li C, et al. Frame-wise action representations for long videos via sequence contrastive learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13801-13810.

[11] Dwibedi D, Aytar Y, Tompson J, et al. Temporal cycle-consistency learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1801-1810.

[12] Dvornik M, Hadji I, Derpanis K G, et al. Drop-dtw: Aligning common signal between sequences while dropping outliers[J]. Advances in Neural Information Processing Systems, 2021, 34: 13782-13793.

[13] Wang X, Zhu L, Wu Y, et al. Symbiotic attention for egocentric action recognition with object-centric alignment[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

[14] Huang C, Wen J, Xu Y, et al. Self-supervised attentive generative adversarial networks for video anomaly detection[J]. IEEE transactions on neural networks and learning systems, 2022.

[15] Chen D, Yue L, Chang X, et al. NM-GAN: Noise-modulated generative adversarial network for video anomaly detection[J]. Pattern Recognition, 2021, 116: 107969.

[16] Wang T, Qiao M, Lin Z, et al. Generative neural networks for anomaly detection in crowded scenes[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(5): 1390-1399.

[17] Wang S, Miao Z. Anomaly detection in crowd scene[C]//IEEE 10th International Conference on Signal Processing Proceedings. IEEE, 2010: 1220-1223.

[18] Ravanbakhsh M, Nabi M, Mousavi H, et al. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 1689-1698.

[19] Luo W, Liu W, Gao S. A revisit of sparse coding based anomaly detection in stacked rnn framework[C]//Proceedings of the IEEE international conference on computer vision. 2017: 341-349.

[20] Lu C, Shi J, Jia J. Abnormal event detection at 150 fps in matlab[C]//Proceedings of the IEEE international conference on computer vision. 2013: 2720-2727.

[21] Ramachandra B, Jones M. Street scene: A new dataset and evaluation protocol for video anomaly detection[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 2569-2578.