Unsupervised Video Anomaly Detection Based on Graph Attention Propagation and Semantic Information

Qinghao Kong, Wanru Xu, Zhenjiang Miao, and Ruizhao Zhai
Beijing Jiaotong University, Beijing, China
23125317@bjtu.edu.cn

Abstract. Video anomaly detection (VAD) represents an indispensable computer vision task that underpins security surveillance and public safety maintenance. Compared with supervised paradigms, unsupervised VAD aligns better with practical deployment scenarios where unknown abnormal events are extremely scarce, yet current approaches are confronted with three prominent limitations: insufficient temporal dependency modeling, inconsistent video semantic understanding, and imprecise fine-grained anomaly localization, all of which collectively cause deviation in final anomaly scoring. To address the above drawbacks, this paper puts forward a novel unsupervised VAD framework that integrates graph attention propagation with multimodal semantic information. Specifically, the framework first fuses video semantic features and motion features to build a dynamic spatiotemporal graph, and optimizes node feature representations through graph attention propagation embedded with orthogonal constraints; afterwards, it divides the input video into semantically consistent event segments via a statistical boundary detection module; finally, it leverages a hierarchical prompting strategy to guide multimodal large language models (MLLMs) to generate event semantic descriptions and initial anomaly scores, and further calibrates these scores through video-text semantic alignment to yield precise frame-level anomaly scores. Experimental validations are conducted on the UCF-Crime and XD-Violence datasets with frame-level AUC as the evaluation metric. Results indicate that the proposed framework reaches state-of-the-art performance under both unsupervised and zero-shot settings, and achieves substantial improvements over existing LLM-based VAD methods and even multiple weakly supervised algorithms, which fully demonstrates its effectiveness and robustness.

Keywords: Video Anomaly Detection, Graph Attention Network, Multimodal Large Language Model.

1 Introduction

Video anomaly detection (VAD) is defined as the task of automatically recognizing events that deviate from regular spatiotemporal behavioral patterns in surveillance footage, which boasts broad application prospects in scenarios including security monitoring, traffic operation management, industrial production inspection, intelligent elderly care, and public security governance. As a typical unsupervised learning task, unsupervised VAD models are trained exclusively on normal video samples, and detect

abnormal events by characterizing the distribution of normal spatiotemporal patterns and quantifying the deviation degree of test samples without relying on any anomaly annotation labels. In contrast to supervised detection methods, this paradigm is more adaptable to real-world scenarios where abnormal events are rare and unpredictable, and exhibits prominent strengths in model generalization capability and deployment cost control. In recent years, driven by the fast advancement of multimodal large models and vision-language pre-training technologies, their powerful semantic comprehension and temporal reasoning capacities have further boosted the open-set generalization, zero-shot detection performance, and fine-grained anomaly localization accuracy under the unsupervised learning framework.

Current LLM-driven video anomaly detection approaches can mostly realize coarse-grained localization of abnormal frames through per-frame semantic understanding. However, large language models generally have inherent defects in temporal dependency modeling, present insufficient consistency in long-sequence video understanding, and perform unsatisfactorily in fine-grained anomaly localization, all of which will further introduce systematic bias to the subsequent anomaly scoring process. Targeting these bottlenecks, this paper designs a novel unsupervised video anomaly detection framework. The framework first fuses semantic features with motion features, obtains event-aware video feature representations through graph attention propagation, and segments the complete video into independent event units using a statistical boundary detection algorithm. Afterwards, a multimodal large language model is adopted to generate semantic descriptions for each segmented event unit, and the video-text matching anomaly scores are calibrated accordingly, so as to output the final anomaly score for subsequent abnormal event judgment.

Ablation experiments and comparative tests are carried out on two public benchmarks, namely UCF-Crime and XD-Violence. Experimental findings verify that the proposed framework achieves state-of-the-art classification performance under both the standard unsupervised paradigm and the LLM-based video anomaly detection technical route.

The main contributions of this work are concluded as follows:

- A novel unsupervised video anomaly detection framework is proposed, which realizes event-level anomaly assessment for video content by combining graph attention propagation and multimodal semantic fusion.
- A graph attention propagation mechanism is elaborately designed to support more precise event boundary detection and event segmentation, and video-text semantic alignment is introduced to optimize the anomaly feature cues and scoring results output by multimodal large language models (MLLMs).
- The proposed framework achieves competitive performance on both UCF-Crime and XD-Violence datasets under zero-shot and unsupervised experimental settings, and its detection effect even exceeds multiple classical weakly supervised detection algorithms.

2 Related Work

2.1 Video Anomaly Detection

The advancement of deep learning techniques has propelled the fast evolution of video anomaly detection research, forming three main technical branches: supervised learning, unsupervised learning, and weakly supervised learning. Supervised detection methods rely heavily on manually annotated data, yet the extremely low occurrence frequency of abnormal events in real videos leads to the difficulty of acquiring high-quality annotated datasets, which severely restricts the landing of such methods in practical scenarios. Semi-supervised learning schemes effectively alleviate this dilemma, whose core logic is to train the detection model only with normal video samples, and judge abnormal events by calculating the feature deviation between test samples and the learned normal pattern distribution. Typical representative algorithms include reconstruction models built upon Autoencoders (AE) and generative models constructed based on Generative Adversarial Networks (GAN). As for weakly supervised learning methods, they only demand video-level anomaly tags without frame-level or segment-level annotation work, and achieve fine-grained temporal localization of abnormal events by adopting strategies like multiple instance learning and self-supervised learning. This paradigm effectively cuts down annotation expenditure while guaranteeing detection precision, and is more applicable for deployment in large-scale real application scenes.

2.2 Unsupervised Video Anomaly Detection

Unsupervised learning algorithms can fulfill anomaly detection without any annotated data support, by excavating the inherent regularities embedded in video data, such as spatiotemporal continuity and motion consistency. The fundamental hypothesis of such methods is that abnormal events only occupy a tiny proportion of the whole video and present substantial feature differences from normal events. Current mainstream unsupervised algorithms mainly include optical flow reconstruction-based models and spatiotemporal cube missing frame prediction-based models.

For instance, Al et al. [1] implemented anomaly detection by clustering the full dataset, and identified samples in minor clusters as abnormal events. Thakare et al. [3] adopted OneClassSVM and iForest algorithms to detect anomaly samples that fall outside the pre-constructed hypersphere boundary. Zaheer et al. [2] put forward a generative cooperative learning approach, which recognizes abnormal events according to the local contrast of video clips, namely the feature difference between adjacent video segments. Such methods can accurately delineate the boundary of abnormal events, but usually cause attenuation of anomaly features inside the event segments.

In terms of LLM-based unsupervised VAD research, LAVAD [7] leverages channel-aligned MLLMs to perform query optimization on the anomaly scores output by large language models (LLMs). On this basis, MCANet [6] proposed a training-free video anomaly detection framework, which dynamically generates multimodal text descriptions and conducts semantic analysis; it integrates vision-language models, audio-

language models and large language models to realize efficient anomaly detection. It is worth noting that normal and abnormal events may coexist in the same surveillance scene, and it is usually challenging to distinguish them effectively without pre-defining the feature distribution of normal samples..

3 Method

For an input video sequence V composed of consecutive frames, the core objective of this work is to accurately detect frames containing abnormal contents. Let $\{a_t\}_{t=1}^{T}$ represent the set of anomaly indicators, where $a_t \in \{0,1\}$ stands for the binary anomaly state corresponding to the t-th frame. The overall architecture of the proposed model is shown in Fig. 1, which comprises three core modules in total. Specifically, the method first builds a dynamic spatiotemporal graph and optimizes node feature representations via graph attention propagation; afterwards, the enhanced feature sequences are fed into a statistical boundary detection module, which identifies event transition points and acquires accurate event boundaries through temporal discontinuity analysis; finally, an event-centric scoring module is designed to assess the anomaly degree inside the detected event units with the help of semantic information, and the scoring results are further optimized via video-text feature alignment.

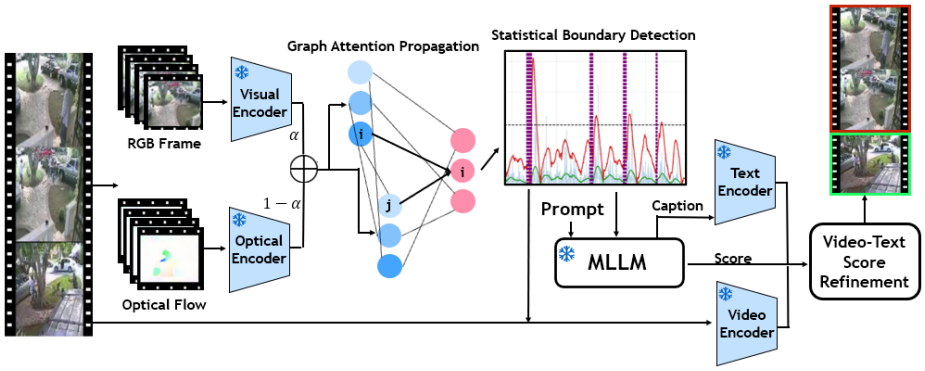

Fig. 1. The overall pipeline of the model is as follows: First, motion and appearance features are extracted to construct a dynamic graph, and graph attention propagation improves node representations through message passing with orthogonal constraints. Subsequently, the statistical boundary detection stage identifies event boundaries by combining the composite divergence metric with an adaptive threshold. Finally, the event-centric scoring stage uses MLLMs to evaluate the anomaly probability within semantically coherent event units, and further refines the anomaly scores using video-text alignment to ultimately obtain the judgment on video anomalies.

3.1 Graph Attention Propagation

To characterize the correlation between different nodes, this work constructs a dynamic graph structure with time decay characteristics. The model adopts a two-stream feature

extraction architecture to integrate semantic patterns and motion patterns, where the semantic branch employs the CLIP model to generate discriminative feature embeddings and performs L2 normalization processing, while the motion branch extracts temporal dynamic features through optical flow calculation. The calculation formula for semantic-motion feature fusion is expressed as follows:

$$f_i = \alpha f_i^{clip} \oplus (1-\alpha) f_i^{flow} \tag{1}$$

where α denotes the semantic-motion fusion coefficient, f_i^{clip} corresponds to the output feature of the semantic branch, f_i^{flow} represents the output feature of the motion branch, and the fused feature vector f_i incorporates both semantic and motion constraints to support precise identification of event transition boundaries.

n the constructed dynamic spatiotemporal graph model G=(F, E), the vertex set F corresponds to video frame units with enhanced feature encoding, and the edge set E is dynamically built based on cross-modal similarity measurement criteria. The time-decay adjacency matrix E is defined as:

$$E_{\mathrm{i,j}} = \frac{\alpha \cdot \cos(f_i^{clip}, f_j^{clip}) + (1-\alpha) \cdot \exp(-\left\| f_i^{flow} - f_j^{flow} \right\|)}{1 + \gamma \cdot \left| i - j \right|} \tag{2}$$

In this formula, the semantic-motion fusion coefficient is set to 0.75, which shares the same value with the node similarity calculation coefficient. To strengthen short-term feature correlation and weaken long-distance feature association, the model integrates the multimodal similarity between frame i and frame j in the numerator part, and introduces a time decay factor \(\gamma\) in the denominator to restrict long-distance node connections. The weight of this factor presents a decreasing trend with the growth of temporal distance between nodes.

Subsequently, a graph attention network embedded with orthogonal constraints is established. For the node features $F^{(0)} = [f_1, ..., f_n]^T$ obtained from the previous propagation step, the model projects them into an orthogonal subspace to avoid dimension collapse, and the specific projection method is shown in Equation (3):

$$\begin{cases} Q = QR\left\{ N(0,1)^{d \times k} \right\} \\ K = QR\left\{ N(0,1)^{d \times k} \right\} \\ V = QR\left\{ N(0,1)^{d \times d} \right\} \end{cases} \tag{3}$$

where QR operation guarantees orthonormal column vectors through QR decomposition, $d$represents the dimension of fused features, and k stands for the dimension of the projected subspace. These fixed orthogonal matrices can maximally retain the original feature information.

For any node f_i\(f_i\), its attention weight corresponding to neighbor node f_j in the dynamic graph is calculated by:

$$Atten_{i,j} = Softmax(\frac{(f_i Q)(f_j K)^T}{\sqrt{d_a}})(f_j^{(t)} V) \tag{4}$$

The iterative update rule of node features is formulated as:

$$f_i^{(t+1)} = f_i^{(t)} + \sum_{j \in N_i} Atten_{i,j} \cdot \mathrm{E}_{(i,j)} \tag{5}$$

With the iterative update of graph attention, the feature propagation process can maintain both global feature divergence and local feature consistency, thus retaining the relative feature differences that play a critical role in anomaly detection tasks.

3.2 Statistical Boundary Detection

On the basis of enhanced features output by the graph attention network, this paper designs a statistical boundary detection mechanism, which recognizes event transition positions by analyzing temporal discontinuity of features. The input video is divided into independent event segments to support higher-quality semantic information generation in subsequent modules.

For two adjacent frames $(i, i+1)$ \((i, i+1)\), the composite divergence metric is calculated by:

$$s_i = \left\| f_{i+1} - f_i \right\|_2^2 + (1 - \frac{f_i^T f_{i+1}}{\left\| f_i \right\| \left\| f_{i+1} \right\|}) \tag{6}$$

On this basis, the moving average of the divergence metric is computed:

$$\mu_i = \frac{1}{w} \sum_{k=i-\lfloor w/2 \rfloor}^{i+\lfloor w/2 \rfloor} s_i \tag{7}$$

Then the signal ratio r_i can be obtained by:

$$r_i = \frac{s_i}{\mu_i} \tag{8}$$

The uniform window weighting strategy in the moving average calculation can suppress instantaneous signal fluctuation and jitter, and improve the robustness of the algorithm against noise. To lower the sensitivity to outlier samples caused by occlusion or other environmental interference, this work adopts the median absolute deviation method to generate an adaptive threshold, and the specific calculation is as follows:

$$\begin{aligned} MAD(r_i) &= median(\left| r_i - median(r_i) \right|) \\ M &= median(r_i) + k \times MAD(r_i) \end{aligned} \tag{9}$$

Finally, a boundary point merging operation is performed on adjacent detected boundaries to avoid over-segmentation problems and further elevate the localization accuracy of event boundaries.

3.3 Video-Text Score Refinement

By elaborately designing a prompting mechanism to unleash the potential of large language models in temporal information aggregation and anomaly score estimation, the general multimodal large models can be converted into efficient video anomaly detection tools.

This work adopts a hierarchical prompting strategy to guide MLLMs in completing two-stage reasoning tasks. In the first reasoning stage, MLLMs generate detailed content descriptions of video events, and extract both explicit surface semantic features and implicit potential semantic features from the descriptions. In the second stage, the model outputs corresponding anomaly scores according to the generated content descriptions, so as to realize cross-modal anomaly assessment and reduce systematic scoring deviation.

The initial anomaly scores are usually generated only based on text subtitle information without incorporating the global feature distribution of the whole dataset. To optimize the initial scoring results, this work performs feature extraction via both text encoders and video encoders, and aggregates the anomaly scores of frames with high semantic similarity to calibrate the initial scores. The specific calculation formula is shown as follows:

$$\tilde{a}_i = \sum_{k \in \mathbf{K}_i} a_k \cdot \frac{e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k) \rangle}}{\sum_{k \in \mathbf{K}_i} e^{\langle \mathcal{E}_V(\mathbf{V}_i), \mathcal{E}_T(S_k) \rangle}} \tag{10}$$

After the video-text score refinement process, the final anomaly score corresponding to each frame in the video can be acquired, which supports subsequent judgment on whether abnormal events exist in the video.

4 Experiments

This section elaborates on the experimental configuration of this work, covering dataset introduction, evaluation metric selection, and implementation detail settings. Afterwards, comparative experiments between the proposed method and state-of-the-art unsupervised VAD algorithms are conducted. Finally, qualitative analysis results and module ablation experimental findings are presented and discussed.

4.1 Experimental Setup

DataSets. Two public datasets extensively adopted in the video anomaly detection domain are selected for performance evaluation, namely UCF-Crime and XD-Violence.

UCF-Crime is a large-scale benchmark dataset widely utilized in recent VAD research, which is characterized by diverse surveillance scenes. The dataset is composed of long untrimmed surveillance videos, covering 13 types of real-world abnormal events that have significant impacts on public security. The complete dataset contains 950 unedited real-world surveillance videos with different anomaly types, and an equal number of normal video samples, reaching a total of 1900 video clips.

XD-Violence is a large-scale audio-visual benchmark specially constructed for video violence behavior detection. It supports both binary classification and multi-label classification tasks, and can be applied in research fields including video anomaly detection and surveillance video content analysis.

Evaluation Metrics. In all experiments of this work, the frame-level Area Under the Receiver Operating Characteristic Curve (AUC), a universally recognized evaluation indicator in this field, is selected as the core measurement standard. It should be clarified that the ROC curve is drawn by adjusting the anomaly score threshold during the inference phase, and a higher AUC value indicates better overall detection performance of the model.

Implementation Details. The feature encoder of the proposed model adopts a dual-branch architecture that integrates visual semantic features and motion representation features. Specifically, the CLIP ViT-B/16 pre-trained model is employed to generate 512-dimensional semantic embedding features, and optical flow features are extracted via the pre-trained RAFT model to obtain 128-dimensional feature vectors. Bilinear interpolation is applied for spatial alignment operation to maintain the consistency of feature resolution. The scoring module is built upon the Video LLaMA2.1-7B-16F model as the backbone network. The main hyperparameters are set as follows: the temporal decay factor is set to 0.6, the semantic-motion fusion coefficient is set to 0.75, and the graph attention propagation process is executed with only one iteration.

4.2 Main Results

To comprehensively assess the detection performance of the proposed method, this paper conducts sufficient comparative experiments with mainstream state-of-the-art algorithms on two widely-used video anomaly detection benchmarks, UCF-Crime and XD-Violence. All comparison methods are evaluated with the frame-level AUC (%) metric, which serves as the standard evaluation criterion in the video anomaly detection field. The specific experimental results are listed in Table 1, and the main conclusions are summarized as follows.

On the UCF-Crime dataset, the proposed method reaches the highest AUC value of 82.91%, outperforming all baseline algorithms involved in the comparison. It achieves substantial performance improvements over traditional detection methods including GCL, LANP and DyAnNet, which verifies the outstanding capability of our method in modeling complex real-world anomaly scenes. When compared with recent ViT-based methods such as LLaVA, Video-Llama2, LAVAD and EventVAD, the proposed method still achieves noticeable performance gains, further confirming the effectiveness of the designed technical framework. On the XD-Violence dataset, the proposed method also maintains state-of-the-art detection performance with an AUC score of

88.11%. This result proves that the method possesses strong generalization capacity for different anomaly categories (such as violent behaviors), and can well adapt to the data characteristics of the XD-Violence dataset.

Table 1. Results on the ucf-crime and xd-violence datasets.

Method	Backbone	UCF-Crime AUC(%)	XD-Violence AUC(%)
GCL	ResNext	71.04	--
DyAnNet	I3D	79.76	--
LANP	ResNext	76.64	--
LLaVA	ViT	72.84	79.62
Video-Llama2	ViT	74.42	80.21
LAVAD	ViT	78.33	82.89
EventVAD	ViT	82.03	87.51
Ours	ViT	**82.91**	**88.11**

Generally speaking, the proposed method achieves optimal detection performance on both experimental datasets, presenting stable performance improvement compared with all baseline methods, especially for the latest advanced ViT-based detection schemes, which fully demonstrates the technical superiority of the proposed framework.

4.3 Ablation Study

The optical flow extraction module, graph attention propagation module, and video-text score alignment module are identified as the three core components that affect the overall detection performance of the model. To quantitatively analyze the contribution of each component to the final performance, this work carries out a modular ablation experiment targeting the three key modules. The experimental results are displayed in Table 2. It can be seen from the data that the introduction of each module can bring significant improvement to the model's detection accuracy, which verifies the effectiveness of all three modules and proves that none of them can be dispensed with.

In addition, to explore the influence of backbone large model selection on detection performance, this study selects large models with different parameter scales and structural types as the backbone network, and conducts comparative experiments on the XD-Violence dataset. The corresponding experimental results are presented in Table 3.

Table 2. Ablation study of different modules.

GANP	RATF	Refinement	AUC(%)
×	×	×	70.21
×	×	✓	78.33
×	✓	×	74.21
✓	×	×	73.78
✓	×	✓	85.69
×	✓	✓	84.88
✓	✓	×	82.26
✓	✓	✓	**88.11**

The experimental findings explicitly show that video-domain specialized large models are much more applicable to video anomaly detection tasks than general-purpose large language models. This conclusion indicates that video-specific large models are equipped with better visual-temporal feature comprehension ability, thus being more suitable for video anomaly detection tasks that involve both visual feature extraction and temporal dynamic modeling. On the contrary, general large language models lack targeted modeling design for video visual features and spatiotemporal dynamic patterns, resulting in apparent performance deficiencies in such tasks.

Table 3. The impact of different basic LLMs on anomaly detection.

Base-model	Parameters	AUC(%)
Llama2-7b	7B	78.44
Llama2-13b	13B	80.25
Video-Llama2	7B	**88.11**

4.4 Qualitative Analysis

To intuitively assess the model's anomaly detection capability and event semantic understanding performance in practical surveillance scenes, this work conducts qualitative analysis on both the UCF-Crime and XD-Violence datasets, and the visualization results are shown in Fig. 2. The figure presents the event segmentation outcomes, the anomaly score curves at frame level generated by the model, and the comparison between these results and the ground truth for both normal and abnormal video samples.

For video samples containing abnormal events, the model can accurately locate the temporal boundaries of abnormal segments, and the generated anomaly score curves are highly consistent with the ground truth annotations. Meanwhile, the event segmentation module can split long video sequences into semantically coherent sub-event units, which supports precise localization and interpretable analysis of abnormal events. For normal scene samples without any abnormal contents, the model consistently keeps

anomaly scores at an extremely low level, which perfectly conforms to the zero-anomaly state of ground truth. This characteristic reflects the model's excellent discriminative ability between normal and abnormal behaviors, and can effectively lower the false alarm rate in practical deployment.

On the whole, the qualitative experimental results sufficiently prove the outstanding performance of the proposed model in multiple dimensions, including anomaly detection accuracy, temporal localization precision, and scene interpretability.

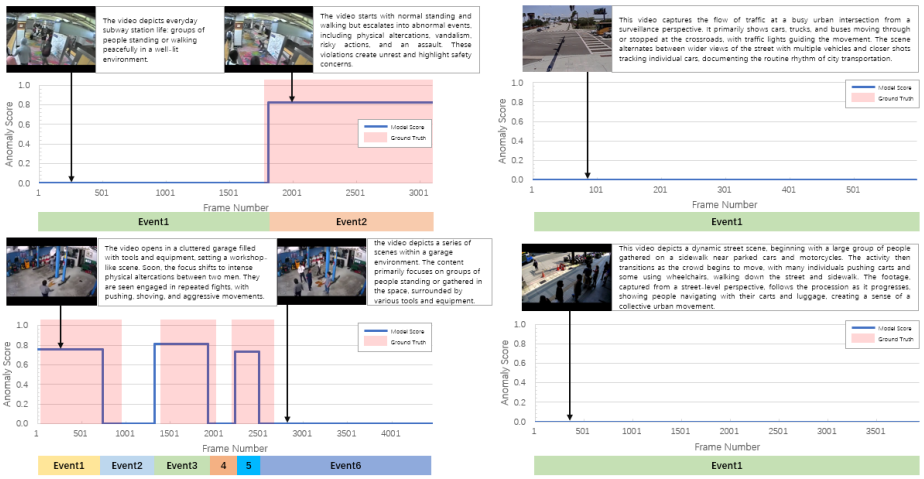

Fig. 2. Visualization of normal and abnormal samples in UCF-Crime and XD-Violence. The figure shows the event segmentation results and the abnormal frame scores of the MLLM, compared with the ground truth.

5 Conclusion

This paper puts forward an unsupervised video anomaly detection framework integrating graph attention propagation and multimodal semantic information, which includes three core technical steps: spatiotemporal feature enhancement based on graph attention mechanism, event segmentation realized by statistical boundary detection, and anomaly score optimization through video-text feature alignment. The framework fuses CLIP-extracted semantic features and RAFT-extracted motion features to build a dynamic spatiotemporal graph, adopts graph attention with orthogonal constraints to optimize node feature representations, divides videos into semantic event units via composite divergence metric calculation and adaptive threshold judgment, and calibrates the initial scores output by MLLMs through video-text semantic alignment to alleviate scoring bias.

For future research directions, this work will focus on introducing audio features to construct multimodal VAD systems, exploring multi-scale graph attention propagation mechanisms, designing adaptive prompting strategies for MLLMs, and optimizing the

framework to support real-time inference, so as to better satisfy the practical application demands of intelligent surveillance systems.

Acknowledgments. This work is supported by National Natural Science Foundation of China 62576028, 62006015, the Beijing Natural Science Foundation 4242028,the Open Fund of State Key Laboratory of Nuclear Power Safety Technology and Equipment(SKL-2025-TS-13) and CEFLA Audio-Video Restoration and Evaluation Key Lab of Ministry of Culture and Tourism.

References

1. Al-Lahham, A., Tastan, N., Zaheer, M.Z., Nandakumar, K.: A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection. In: WACV. pp. 6793–6802 (2024)
2. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: CVPR. pp. 14744–14754 (2022)
3. Thakare, K.V., Raghuwanshi, Y., Dogra, D.P., Choi, H., Kim, I.J.: Dyannet: A scene dynamicity guided self-trained video anomaly detection network. In: WACV. pp. 5541–5550 (2023)
4. Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In CVPR.
5. 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 ECCV.
6. Prabhu Prasad Dev, Raju Hazari, and Pranesh Das. 2025. MCANet: Multimodal Caption Aware Training-Free Video Anomaly Detection via Large Language Model. In International Conference on Pattern Recognition. Springer, 362–37.
7. Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. 2024. Harnessing large language models for training-free video anomaly detection. In Proceedings ofthe IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18527–18536.
8. 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.
9. 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.
10. Zaheer M Z, Mahmood A, Khan M H, et al. Generative cooperative learning for unsupervised video anomaly detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 14744-14754.
11. Wu P, Liu J, Shi Y, et al. Not only look, but also listen: Learning multimodal violence detection under weak supervision[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 322-339.
12. Chen Y, Liu Z, Zhang B, et al. Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2023, 37(1): 387-395.
13. Shi H, Wang L, Zhou S, et al. Learning anomalies with normality prior for unsupervised video anomaly detection[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 163-180.
14. Wang Z, Zou Y, Zhang Z. Cluster attention contrast for video anomaly detection[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 2463-2471.

15. Wu P, Liu J. Learning causal temporal relation and feature discrimination for anomaly detection[J]. IEEE Transactions on Image Processing, 2021, 30: 3513-3527.
16. Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6479-6488.
17. Hasan M, Choi J, Neumann J, et al. Learning temporal regularity in video sequences[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 733-742.
18. Al-Lahham A, Tastan N, Zaheer M Z, et al. A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 6793-6802.
19. Cho M A, Kim M, Hwang S, et al. Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 12137-12146.
20. Yang Z, Guo Y, Wang J, et al. Towards video anomaly detection in the real world: A binarization embedded weakly-supervised network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(5): 4135-4140.
21. Fan Y, Wen G, Li D, et al. Video anomaly detection and localization via gaussian mixture fully convolutional variational autoencoder[J]. Computer Vision and Image Understanding, 2020, 195: 102920.
22. Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 26296-26306.
23. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.
24. Tudor Ionescu R, Smeureanu S, Alexe B, et al. Unmasking the abnormal events in video[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2895-2903.
25. Liu Y, Li C L, Póczos B. Classifier two sample test for video anomaly detections[C]//BMVC. 2018: 71.
26. Hinami R, Mei T, Satoh S. Joint detection and recounting of abnormal events by learning deep generic knowledge[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3619-3627.
27. Zaheer M Z, Mahmood A, Astrid M, et al. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 358-376.
28. Wu P, Liu J, Shen F. A deep one-class neural network for anomalous event detection in complex scenes[J]. IEEE transactions on neural networks and learning systems, 2019, 31(7): 2609-2622.
29. Zhang C, Li G, Qi Y, et al. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 16271-16280.
30. Zhou H, Yu J, Yang W. Dual memory units with uncertainty regulation for weakly supervised video anomaly detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(3): 3769-3777.
31. Wang J, Cherian A. Gods: Generalized one-class discriminative subspaces for anomaly detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8201-8211.

32. Liu Z, Nie Y, Long C, et al. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13588-13597.
33. Yu G, Wang S, Cai Z, et al. Cloze test helps: Effective video anomaly detection via learning to complete video events[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 583-591.
34. Thakare K V, Dogra D P, Choi H, et al. Rareanom: A benchmark video dataset for rare type anomalies[J]. Pattern Recognition, 2023, 140: 109567.
35. Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(7): 971-987.
36. Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.
37. Chan E R, Lin C Z, Chan M A, et al. Efficient geometry-aware 3d generative adversarial networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16123-16133.
38. Gupta I, Kumari S, Jha P, et al. Leveraging LSTM and GAN for modern malware detection[J]. arXiv preprint arXiv:2405.04373, 2024.
39. Wang W, Lv Q, Yu W, et al. Cogvlm: Visual expert for pretrained language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 121475-121499.
40. Yan C, Zhang S, Liu Y, et al. Feature prediction diffusion model for video anomaly detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 5527-5537.