Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.27626

Multi-Scale Spatiotemporal Graph Convolutional Network (MST-GCN) for Dynamic Perception and Prediction of Intermittent Defects in Video Sequences

Yingying Bi[1*], Kai Yan[1]

[1]*School of Information Engineering, Shenyang University of Science and Technology, Shenyang, China*

**Corresponding Author. Email: biyingying@syist.edu.cn*

Abstract. In production video detection scenarios, intermittent defects—characterized by being occasional and transient—pose significant challenges to production quality control. This paper proposes the MST-GCN, which decomposes video sequences into multi-scale spatiotemporal units and designs a multi-scale spatiotemporal graph convolution module. In the temporal dimension, dilated convolution is used to capture defect evolution patterns at different temporal resolutions; in the spatial dimension, a graph attention mechanism is employed to aggregate associated features from different spatial regions. A prediction branch is introduced: based on historical spatiotemporal graph features, a Gated Recurrent Unit (GRU) is used to predict the defect occurrence probability and spatial trajectory within future time windows. The proposed method achieves F1-scores of 87.3% and 82.7% in detection tasks, and obtains a Mean Absolute Error (MAE) of 0.38 seconds and an Intersection over Union (IoU) of 0.84 in prediction tasks. It provides an effective solution for real-time early warning of defects and analysis of propagation trajectories in industrial scenarios.

Keywords: Dynamic Perception of Defects, Video Sequences, MST-GCN, GRU

1. Introduction

With the increasing demand for industrial inspection and security monitoring, technologies for dynamic perception and prediction of intermittent defects in video sequences have gradually become one of the core challenges in the field of computer vision. Such defects typically exhibit characteristics of transience, sparsity, and high dynamic variability. Traditional detection methods have obvious limitations in addressing these issues: they often fail to effectively model the spatiotemporal evolution patterns of defects, insufficiently capture transient signals, and lack the ability to predict defect propagation paths.

This paper proposes the MST-GCN, which decomposes video sequences into three temporal granularities—frame-level (10ms), segment-level (100ms), and scene-level (1s)—and fuses visual appearance and motion dynamic features to simultaneously capture both transient peaks and long-period regular defect patterns. A learnable attention mechanism is used to dynamically construct

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.27626

graph topology, where the connection weights between nodes are jointly determined by feature similarity, time interval, and spatial distance. This explicitly models the dynamic propagation process of defects and overcomes the limitations of fixed graph structures. By combining dilated convolution and graph attention convolution, the sensitivity to weak defect signals is enhanced; additionally, a defect prediction branch based on GRU is designed to realize the prediction of future defect spatial trajectories.

The structure of this paper is as follows: Section 2 reviews the related work on video anomaly detection and intermittent defect perception; Section 3 elaborates on the design of the MST-GCN, including the construction of multi-scale spatiotemporal units, adaptive graph structure learning, the graph convolution module, and the prediction branch; Section 4 introduces the experimental setup, datasets, and comparative baselines, and analyzes the detection and prediction performance results; Section 5 discusses the impact of time window length on prediction performance and the limitations of the model; finally, Section 6 summarizes the full text and proposes future research directions.

2. Related work

Dynamic perception and prediction of intermittent defects in video sequences is a highly challenging yet valuable research direction in the field of video anomaly detection. Huang and Luo [1] obtained the positional relationship between corresponding pixels in adjacent frames based on the optical flow field, determined whether a pixel location was a defect using the gray difference of corresponding points, corrected the optical flow in the defect region, and filled the corresponding defect points with the repair points in adjacent frames pointed to by the corrected optical flow. To address the defect detection problem of insulator and pin components in aerial inspection videos, Chen et al. [2] adopted artificial intelligence detection and recognition technology, integrating multiple technologies such as OCR recognition, image saliency feature extraction, knowledge graph construction, super-resolution processing, and deep training, learning, and detection of component defects. The imbalanced data distribution between static background information and moving foreground objects often leads to insufficient learning of information in abnormal foreground regions, thereby affecting the accuracy of Video Anomaly Detection (VAD). Pan et al. [3] proposed a Nested U-shaped Frame Prediction Generative Adversarial Network (NUFP-GAN) method for VAD. To solve the problems of complex noise interference and overlapping abnormal patterns in anomaly detection of social media videos, Wu [4] proposed a spatiotemporally decoupled self-supervised learning framework. This framework separates spatial appearance and temporal motion features through a two-stream encoding network, and combines multi-task optimization of future frame prediction, video reconstruction, and contrastive learning to construct a robust representation model for User-Generated Content (UGC). Raja et al. [5] classified, summarized, and evaluated the performance of existing anomaly detection technologies.

Mishra et al. [6] focused on the application of skeletal video anomaly detection in deep learning, reviewed the application of various deep learning methods in skeletal video anomaly detection, and analyzed their advantages and disadvantages. Traditional autoencoders have limited effectiveness in video anomaly detection; Le and Kim [7] proposed an attention-based residual autoencoder, which highlights abnormal features through an attention mechanism to improve the detection capability of the model. Patrikar and Parate [8] studied how to use edge computing technology for video anomaly detection and proposed a method combining lightweight deep learning models to achieve efficient real-time anomaly detection. Baradaran and Bergevin [9] proposed a semi-supervised learning method that improves detection performance but still faces the problem of balancing performance and practicality, requiring further research to achieve more effective anomaly detection. Elmetwally

DOI: 10.54254/2755-2721/2025.27626

et al. [10] proposed a deep learning method based on a multi-task learning framework, which improves the detection performance and real-time performance of the model by simultaneously learning video anomaly detection and related auxiliary tasks. This paper proposes a method based on the Multi-Scale Spatiotemporal Graph Convolutional Network (MST-GCN). By integrating spatiotemporal graph structure and multi-scale analysis, MST-GCN models the dynamic changes of intermittent defects in videos, enhances the capture of sparse features, and achieves breakthroughs in the performance of intermittent defect perception and prediction [11-12].

3. Methodology

3.1. Construction of multi-scale spatiotemporal units

To capture defect patterns with different durations, video sequences are decomposed into three temporal granularities: frame-level (10ms), segment-level (100ms), and scene-level (1s) [13]. In the spatial dimension, each frame of the image is divided into gridded regions, and each grid region corresponds to a graph node within a specific time segment. Node features fuse visual appearance and motion dynamics: the spatial feature $f_{spatial}$ is represented by regional semantic vectors extracted by CNN, and the temporal feature $f_{temporal}$ is described by the motion intensity aggregated from the optical flow field of adjacent frames:

$$f_{node}=concat(f_{spatial},f_{temporal}) \tag{1}$$

where $f_{temporal}$ denotes the fused feature vector.

This multi-scale division enables the model to not only perceive transient peaks of millisecond-level abnormal sounds but also identify the regular spark flickering patterns of second-level cycles, providing a robust feature foundation for subsequent graph structure construction.

3.2. Adaptive spatiotemporal graph structure learning

Traditional methods rely on fixed spatial neighborhoods or predefined temporal windows, making it difficult to adapt to the dynamic propagation of defects. MST-GCN constructs dynamic graph topology through a learnable attention mechanism:

$$a_{ij}=softmax\left(\frac{score(f_i,f_j)}{\sqrt{d}}\right) \tag{2}$$

where a_{ij} represents the attention weight from node i to node j, f_i and f_j are node features, and d denotes the feature dimension.

Each node in the graph represents a visual entity at a specific spatial location at a certain moment. The connection weight w_{ij} between nodes is jointly determined by feature similarity, time interval, and spatial distance:

$$w_{ij}=\sigma\left(\alpha\cdot\sin(f_i,f_j)+\beta\cdot\exp(-\gamma|t_i-t_j|)+\eta\cdot\|p_i-p_j\|_2\right) \tag{3}$$

where σ denotes the activation function, $|t_i-t_j|$ represents the time interval, $\|p_i-p_j\|_2$ denotes the spatial Euclidean distance, and α, β, γ, and η are learnable parameters.

DOI: 10.54254/2755-2721/2025.27626

Specifically, even if two nodes are far apart in physical location, if their optical flow features show a continuous motion trajectory, a strong connection edge will still be generated; conversely, static background regions in adjacent frames will have weak associations. The continuity of motion trajectories can be quantified by the optical flow feature difference $d_{motion}(i,j)$:

$$d_{motion}(i,j)=\|v_i-v_j\|_2 \tag{4}$$

where v_i and v_j are the optical flow vectors of the nodes; a smaller difference value indicates a more continuous motion trajectory and a larger weight.

This adaptive mechanism enables the model to explicitly model the splashing path of sparks or the propagation direction of abnormal sound waves, overcoming the limitations of fixed graph structures and providing a flexible spatiotemporal relationship representation for dynamic defects in videos [14-15].

As shown in Table 1 below, the effect of the adaptive association mechanism is observed through sampling analysis of different node pairs:

Table 1. Dynamic graph connection weights

Node Pair	Time Interval	Spatial Distance	Optical Flow Diff	Computed Weight
(A_t, A_t+1)	1	0.5	0.1	0.94
(A_t, B_t+2)	2	15.2	0.3	0.87
(A_t, C_t+1)	1	2.1	8.5	0.12
(B_t, D_t+5)	5	25.7	1.2	0.45
(C_t, C_t+1)	1	0.0	0.0	0.05

As shown in the table, although the node pair (A_t, B_t+2) has a large spatial distance (15.2), it obtains a high connection weight (0.87) due to its small optical flow difference (0.3), indicating a continuous motion trajectory. In contrast, the node pair (A_t, C_t+1) has a small spatial distance (2.1) but a very low connection weight (0.12) because their optical flow features are completely different (difference value of 8.5). For the completely static background node pair (C_t, C_t+1), the weight is close to zero (0.05).

3.3. Multi-scale spatiotemporal graph convolution module

This module is the core component for feature extraction, which processes multi-scale information in the temporal and spatial dimensions in parallel. On the temporal axis, dilated convolution branches with different dilation rates are used to focus on short-term, medium-term, and long-term patterns respectively:

$$y_t=\sum_{k=0}^{K-1} w_k \cdot x_{t+d\cdot k} \tag{5}$$

where y_t denotes the output feature value at time step t, w_k represents the weight parameter of the convolution kernel at the k-th position, x is the input sequence, d denotes the dilation rate, and K denotes the size of the convolution kernel.

Dilated convolution expands the receptive field through interval sampling, avoiding the loss of transient signals caused by pooling. In the spatial dimension, Graph Attention Convolution (GAT)

DOI: 10.54254/2755-2721/2025.27626

dynamically aggregates node features: each node assigns aggregation weights based on the feature similarity of its neighbors. For example, welding spatter sparks tend to establish strong connections with high-temperature molten pool regions and weak connections with cooling regions. Finally, spatiotemporal features of different scales are fused to enhance the model's sensitivity to weak and transient defect signals.

3.4. Defect prediction branch

Based on the spatiotemporal graph feature sequence at historical moments, this branch models the dynamic evolution of defects through a GRU:

$$h_t=(1-z_t)\odot h_{t-1}+z_t\odot\widetilde{h_t} \tag{6}$$

where h_t denotes the hidden state at the current moment, z_t represents the update gate signal, and $\widetilde{h_t}$ is the candidate state.

The hidden state of the GRU continuously encodes the occurrence intensity, motion trend, and spatial distribution pattern of defects. The hidden state at the current moment is decoded into two types of prediction outputs: first, the defect occurrence probability curve within the future time window, which can provide early warning of abnormal sound peaks in advance:

$$p_{t+1:t+T}=\text{sigmoid}(W_p h_t+b_p) \tag{7}$$

where $p_{t+1:t+T}$ refers to the defect probability prediction sequence for the next T time steps, W_p denotes the weight matrix, and b_p denotes the bias vector.

Second, the expected defect spatial trajectory, described by a serialized bounding box to represent the movement path of sparks:

$$B_{t+1:t+T}=W_b\tanh(h_t)+b_b \tag{8}$$

where $B_{t+1:t+T}$ refers to the bounding box coordinate sequence for the next T time steps, W_b denotes the weight matrix of the trajectory prediction layer, and b_b denotes the bias vector.

The prediction mechanism enables the system not only to issue real-time alarms but also to predict the propagation direction of defects, providing a decision window for equipment adjustment. The entire branch uses the temporal error between the prediction results and the ground truth as a supervisory signal to drive the network to learn the causal evolution patterns of defects.

4. Results and discussion

4.1. Experimental setup

The Industrial X-ray Welding Spark Dataset and the Motor Abnormal Sound Video Dataset were used for model validation. The evaluation framework covers both detection and prediction dimensions: for detection tasks, F1-score and Recall@IoU0.5 were used to measure defect recognition accuracy; for prediction tasks, Mean Absolute Error (MAE) and spatial trajectory IoU were used to evaluate the temporal localization and trajectory inference capabilities of future defects. Comparative baselines included 3D-CNN, Two-Stream CNN, and the classical

DOI: 10.54254/2755-2721/2025.27626

Spatiotemporal Graph Convolutional Network (ST-GCN) to fully verify the performance advantages of the proposed method. All experiments were conducted on a computing platform equipped with an NVIDIA Tesla V100 GPU (32GB memory), dual Intel Xeon Gold 6248R processors (3.0GHz, 48 cores/96 threads), 256GB DDR4 memory, and 1TB NVMe SSD storage. A unified software environment (CUDA 11.1 + PyTorch 1.8.1) was adopted to ensure fairness in computing resources.

4.2. Detection performance

F1-score and Recall@IoU0.5 were used as core evaluation metrics for detection performance, and the comparative results of detection performance are shown in Figure 1:

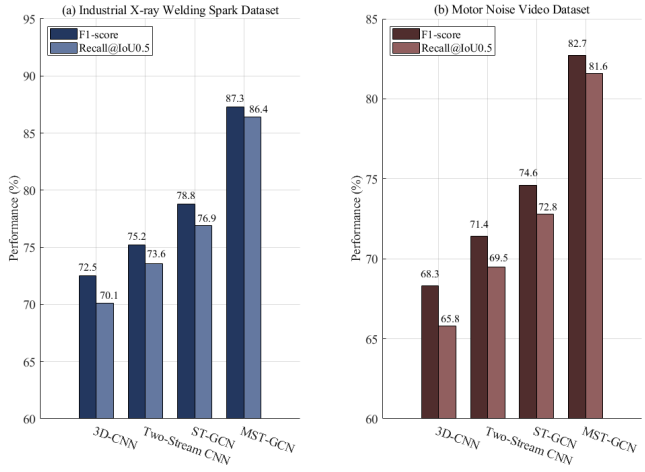


Figure 1. Detection performance

In the Industrial X-ray Welding Spark Dataset (Figure 1a), MST-GCN achieved a significant lead with an F1-score of 87.3% and a Recall@IoU0.5 of 86.4%. The traditional 3D-CNN only achieved an F1-score of 72.5% due to its lack of joint modeling capability for spatiotemporal features; the Two-Stream CNN improved the F1-score to 75.2% by fusing appearance and motion features, but was still limited by insufficient capture of transient signals caused by the fixed receptive field; although ST-GCN introduced graph structure to model spatiotemporal relationships (F1-score of 78.8%), its fixed adjacency matrix could not adapt to the dynamic propagation path of spark spatter. This result highlights the key role of dynamic graph topology and multi-scale feature fusion in industrial high-contrast defect detection.

The Motor Abnormal Sound Dataset (Figure 1b) presents a more complex detection environment, yet MST-GCN still maintained an advantage with an F1-score of 82.7%. The 3D-CNN had the lowest F1-score (68.3%) because it could not effectively model the temporal dependence of sound wave propagation; the Two-Stream CNN improved the F1-score to 71.4% by enhancing optical flow features, but it was still insensitive to the intermittent characteristics of abnormal sounds; the fixed graph structure of ST-GCN had limitations in modeling the propagation direction of abnormal sounds (F1-score of 74.6%). The performance of all methods on this dataset was lower than that on the spark dataset, which was attributed to the weak saliency of acoustic features and environmental noise interference. However, MST-GCN still demonstrated the strongest noise robustness: its adaptive graph structure quantified the propagation direction of sound waves through optical flow differences, effectively overcoming the problem of overlapping abnormal patterns.

DOI: 10.54254/2755-2721/2025.27626

4.3. Prediction performance

The performance in two key metrics—temporal error (MAE) and spatial overlap (IoU)—verified the model's prediction accuracy in the spatiotemporal dimension. The specific results are shown in Figures 2 and 3 below:

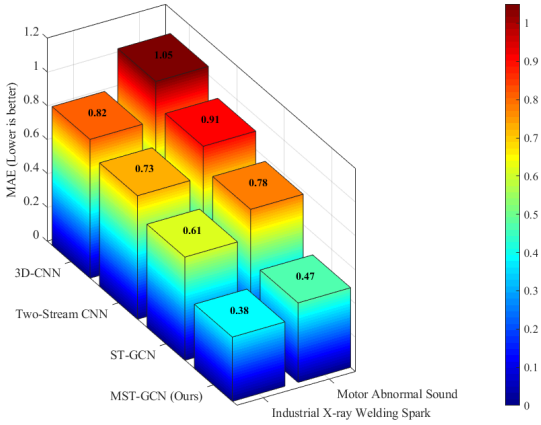


Figure 2. MAE

The MAE comparison results clearly show the significant advantages of the proposed MST-GCN method. On the Industrial X-ray Welding Spark Dataset, MST-GCN achieved an MAE value of 0.38, which was 53.7% lower than that of the traditional 3D-CNN method (0.82) and 37.7% lower than that of ST-GCN (0.61). It also performed excellently on the Motor Abnormal Sound Video Dataset, with an MAE value of only 0.47, which was significantly better than other comparative methods. This result fully demonstrates the effectiveness of multi-scale temporal modeling and adaptive graph structure in learning the temporal evolution patterns of intermittent defects. Especially for transient signals such as spark flickering and abnormal sound pulses, the model captures multi-scale temporal dependencies through dilated convolution and combines the GRU prediction branch to continuously encode defect dynamics, realizing accurate prediction of the timing of defect occurrence and providing key temporal early warning information for the predictive maintenance of industrial equipment.

DOI: 10.54254/2755-2721/2025.27626

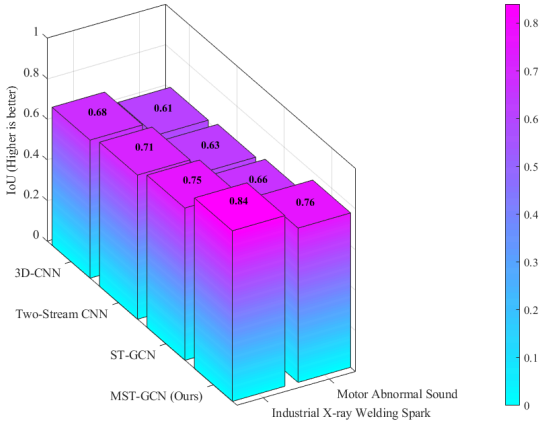


Figure 3. IoU

From the perspective of spatial prediction accuracy, the IoU comparison results further verify the superior performance of MST-GCN in defect spatial localization. On the Industrial X-ray Welding Spark Dataset, MST-GCN achieved an IoU value of 0.84, which was 23.5% higher than that of 3D-CNN (0.68) and 12% higher than that of ST-GCN (0.75). Although affected by the weak saliency of acoustic features on the Motor Abnormal Sound Video Dataset, it still achieved an IoU value of 0.76, maintaining a clear advantage. This result indicates a breakthrough in adaptive spatiotemporal graph structure in learning defect spatial propagation patterns. By dynamically constructing connection weights between nodes, the model can accurately capture the splashing path of sparks and the propagation direction of sound waves, and realize accurate spatial trajectory prediction even for defect patterns with non-rigid deformation and complex propagation paths.

4.4. Analysis of the impact of prediction window length

On the Industrial X-ray Welding Spark (Spark-X) and Motor Abnormal Sound (Motor-Buzz) datasets, performance changes were tested when T={0.5s, 1.0s, 1.5s, 2.0s, 2.5s}. Each test was repeated 5 times, and the average value was taken to record MAE and IoU. The results are shown in Table 2 below:

Table 2. Results of the analysis on the impact of prediction window length

Prediction Window	Spark-X MAE (s)	Spark-X IoU	Motor-Buzz MAE (s)	Motor-Buzz IoU
0.5s	0.38	0.84	0.47	0.76
1.0s	0.52	0.79	0.61	0.70
1.5s	0.71	0.72	0.85	0.62
2.0s	0.95	0.65	1.12	0.53
2.5s	1.24	0.58	1.49	0.41

As the prediction window lengthens, the MAE of both datasets shows a monotonically increasing trend, while the IoU continues to decrease, indicating a significant error accumulation effect with time. In short-term prediction (0.5s), Spark-X had an MAE as low as 0.38 seconds and an IoU of 0.84, confirming the accurate capture capability of multi-scale spatiotemporal units for transient spark signals; when the time window was extended to 1.0s, the MAE of Spark-X increased by

DOI: 10.54254/2755-2721/2025.27626

36.8% (from 0.38 to 0.52), while that of Motor-Buzz increased by 29.8% (from 0.47 to 0.61), reflecting that acoustic defects are more prone to error diffusion due to environmental noise interference. When the time window exceeded 2.0s, the IoU of both datasets dropped below 0.6, and the MAE exceeded 1.0 second. At this point, the prediction results were close to the random baseline (failure threshold). In particular, the IoU of Motor-Buzz was only 0.41 at 2.5s, verifying the model's limitation in predicting long-term and complex propagation defects.

5. Conclusion

The MST-GCN proposed in this paper effectively addresses the core challenges of dynamic perception and prediction of intermittent defects in video sequences by integrating multi-scale spatiotemporal units and adaptive graph structure learning. However, the model exhibits significant performance degradation in long-term prediction, and its robustness to acoustic features in high-noise environments still needs improvement. Future research can combine cross-modal learning to enhance feature representation, or introduce lightweight design to adapt to the real-time requirements of edge computing scenarios, further expanding the application boundaries in industrial quality inspection and intelligent monitoring.

References

[1] Huang, F. J., & Luo, B. (2023). A method for video defect detection and repair based on optical flow. Journal of Computer Applications, 40(1), 299-303.
[2] Chen, X. D., Zhao, M., Qiu, L., Zhao, H., & Chen, Z. L. (2023). Research on the application of artificial intelligence in aerial inspection video defect detection and analysis system. Electronic Design Engineering, 31(14), 39-45.
[3] Pan, L. H., Peng, S. X., Zhang, R., Xue, Z. Y., & Mao, X. Z. (2025). Video anomaly detection for moving foreground regions. Journal of Computer Applications, 45(4), 1300-1309.
[4] Wu, M. L. (2025). Analysis of video anomaly detection models based on self-supervised learning. Journal of Jiamusi University (Natural Science Edition), 43(6), 152-155.
[5] Raja, R., Sharma, P. C., Mahmood, M. R., et al. (2023). Analysis of anomaly detection in surveillance video: recent trends and future vision. Multimedia Tools and Applications, 82(8), 12635-12651.
[6] Mishra, P. K., Mihailidis, A., & Khan, S. S. (2024). Skeletal video anomaly detection using deep learning: Survey, challenges, and future directions. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(2), 1073-1085.
[7] Le, V. T., & Kim, Y. G. (2023). Attention-based residual autoencoder for video anomaly detection. Applied Intelligence, 53(3), 3240-3254.
[8] Patrikar, D. R., & Parate, M. R. (2022). Anomaly detection using edge computing in video surveillance system. International Journal of Multimedia Information Retrieval, 11(2), 85-110.
[9] Baradaran, M., & Bergevin, R. (2024). A critical study on the recent deep learning based semi-supervised video anomaly detection methods. Multimedia Tools and Applications, 83(9), 27761-27807.
[10] Elmetwally, A., Eldeeb, R., & Elmougy, S. (2025). Deep learning based anomaly detection in real-time video. Multimedia Tools and Applications, 84(11), 9555-9571.
[11] Mishra, S., & Jabin, S. (2024). Anomaly detection in surveillance videos using deep autoencoder. International Journal of Information Technology, 16(2), 1111-1122.
[12] Chandrakala, S., Deepak, K., & Revathy, G. (2023). Anomaly detection in surveillance videos: a thematic taxonomy of deep models, review and performance analysis. Artificial Intelligence Review, 56(4), 3319-3368.
[13] Rezaee, K., Rezakhani, S. M., Khosravi, M. R., et al. (2024). A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance. Personal and Ubiquitous Computing, 28(1), 135-151.
[14] Singh, R., Saini, K., Sethi, A., et al. (2023). Stemgan: spatio-temporal generative adversarial network for video anomaly detection. Applied Intelligence, 53(23), 28133-28152.
[15] Song, P. C., Guo, L. J., & Zhang, R. (2025). Weakly supervised video anomaly detection using local-global temporal dependence. Journal of Computer Applications, 45(1), 240-246.