Data Engineering and Applications

Research Paper

GRCNN: A Gated Recurrent Convolutional Neural Network for Robust Temporal-Spatial Video Anomaly Detection

Matěj Černý[1], Adam Hájek[1] and Adéla Černá[1,*]

[1] Institute of Computer Science, Masaryk University, 602 00 Brno, Czech Republic
*Corresponding author: adela.c@ics.muni.cz

Abstract. With the advancement of intelligent monitoring technology, detecting anomalous events in video streams with rich spatiotemporal correlations and dynamic backgrounds has become increasingly difficult. To address the aforementioned issues, this paper proposes a new Gated Recurrent Convolutional Neural Network (GRCNN). The network features a graph-based attention mechanism, dual-branch recursive paths, and dynamic time gating. Through end-to-end training, data augmentation, and regularization, the improved model enhances the sensitivity and noise resistance of the original model. Experiments on the UCSD Ped2, Avenue, and ShanghaiTech datasets show that GRCNN achieved an average AUC of 95.8%, an F1-score exceeding 92%, and an average detection delay reduced to 71 milliseconds. Due to its excellent design, it outperforms the current best baseline in cross-domain transfer and multimodal fusion. According to the experiments, the model remains effective under uncertain conditions. The results indicate that GRCNN has good adaptability and efficiency, making it suitable for real-time monitoring in public safety and other intelligent surveillance systems.

Keywords: *Pattern Recognition, Video Anomaly Detection, Gated Recurrent Network, Temporal-Spatial Feature Fusion, Attention Mechanism, Multi-Modal Learning*

Received on 18 June 2025, Accepted on 17 November 2025, Published on 5 Jan2026

Introduction

With the widespread adoption of video surveillance systems, new concepts of public safety, intelligent transportation, and smart cities have begun to emerge [1]. The continuous development of high-definition cameras and distributed sensor networks has led to a rapid increase in the volume of visual data, significantly raising the demand for automatic anomaly detection in dynamic and heterogeneous environments [2]. Abnormalities must be accurately and promptly identified to prevent harm and optimally utilize resources in public domains (such as public places, factories, and transportation) [3]. Computer vision methods based on background modeling, trajectory extraction, or handcrafted features are often affected by inherent variations in real video streams, such as changes in lighting, crowded scenes, occlusions, and diverse activity patterns [4]. Not suitable for anomaly detection in practical applications, as they are inherently unflexible and insensitive to spatial or temporal fluctuations [5].

Deep learning-based representation learning frameworks surpass previous frameworks in terms of scalability and accuracy [6]. Convolutional Neural Networks (CNNs) are very suitable for extracting multi-level spatial features, while Recurrent Neural Networks (RNNs) can model the sequential and temporal dependencies in video signals [7]. Most existing methods handle spatial and temporal aspects separately, performing poorly when anomalies occur due to subtle spatiotemporal correlations in complex environments [8]. Although some progress has been made, these methods still have room for improvement in handling space and time. Some studies use hybrid architectures to combine spatial and temporal aspects, but there are still some issues, such as information bottlenecks, suboptimal feature fusion, and lack of interpretability [9]. Given the persistent gaps, researchers are striving to develop more unified and expressive neural networks. These neural networks are

capable of learning and inferring comprehensive spatiotemporal features from extended video data sequences [10].

This paper proposes a new framework based on Gated Recurrent Convolutional Neural Networks (GRCNN) that combines the advantages of spatiotemporal feature learning for identifying anomalous events in videos. To address the previous shortcomings in integration and interpretability, the new technology employs dynamic memory and deep spatial representation. The model is suitable for various environments and types of anomalies; it focuses on feature propagation and information flow, achieving improved generalization without a significant increase in computational load. This study not only makes new contributions to theoretical understanding but also provides practical support through extensive empirical research on many public datasets and in-depth comparisons with currently known best methods.

Related Works and Fundamental Theories

Spatiotemporal Representation in Video Analysis

A good spatiotemporal representation is the foundation of video analysis. In recent years, deep learning methods such as convolutional neural networks and recurrent neural networks have been proposed to improve this representation [11]. Traditional descriptors can only collect static and dynamic cues and are easily affected by changes in scenes, scales, and expressions. Deep learning now uses CNNs to extract spatial patterns and RNNs to leverage temporal correlations [12]. 3D convolutional models are used to handle local spatiotemporal regions. In the past, dual-stream architectures separated optical flow and appearance processing tasks, but now they are integrated [13]. This method has issues with high computational demands and insufficient spatiotemporal fusion. To enhance the contextual awareness and generalization ability of video understanding, an attention-centered and memory-augmented architecture was designed [14]. Temporal memory, attention mechanisms, and graph networks are also widely used in many other large-scale, real-world environments [15].

GRCNN and Hybrid Architectures

Gated Recurrent Convolutional Neural Networks (GRCNNs) are a relatively new model. Combined with CNNs and RNNs; by adding gating mechanisms to the convolutional structure, it can selectively store and transmit information in both time and space [16]. To address the long-term dependency problem in video event streams, other hybrid models have also been introduced in addition to GRCNNs. These models combine convolution, recursion, attention, pooling, and graph structures [17]. For example, hybrid models with self-attention and relational reasoning can flexibly focus on important parts of the content and handle background noise or scene changes more robustly [18]. This structure demonstrates improved anomaly detection and generalization capabilities in challenging situations. Its increasingly complex structure requires strict standards and provides new domain adaptation strategies for convergence and preventing overfitting [19]. Due to the demand for interpretable deep learning, these hybrid networks are developing explanation modules to enhance the traceability, transparency, and user trust of critical applications [20].

Key Challenges in Anomaly Event Detection

Although anomalous events are extremely rare and diverse, in most real-world environments, normal patterns are the majority [21]. This imbalance not only makes fully supervised methods less effective but also makes it harder for models to generalize, especially in cross-scenario or cross-domain applications [22]. Many anomalies are merely minor spatiotemporal deviations, which require highly discriminative representation learning and efficient data representation learning to cope with occlusions, lighting changes, and complex dynamics [23]. Real-time operation requirements limit the feasibility of heavy architectures. The detection pipeline must meet robustness and speed requirements [24]. Due to the significant risks and real-world consequences that false positives or missed detections can bring, achieving explainability, transparency, and user trust in monitoring and safety-critical environments is crucial [25].

https://doi.org/10.64972/dea.2026.v5i1.1657d:87-100

Proposed GRCNN-Based Model

Architectural Overview

The GRCNN introduced here addresses the problem of anomaly detection in dynamic and unbounded video data by altering spatial encoding and temporal context retention. Convolutional representation learning and gated memory are two components of this architecture, which are connected to form an integrated module for contextual processing of each frame in both time and space. By using a hierarchical convolutional encoder to map each input frame sequence to a tensor with multiple scales and high-dimensional features, it can be represented as:

$$\mathbf{F}_t^{(0)} = \mathrm{Conv}_{\Omega_0}(\mathbf{X}_t) \qquad \text{Eq.(1)}$$

where $\mathbf{X}_t$ denotes the normalized video input at time t, and $\mathrm{Conv}_{\boldsymbol{\Omega}_0}$ is a parameterized spatial kernel bank optimized for local and global pattern extraction.

Due to its recursive gated propagation mechanism, GRCNN integrates spatial features and temporal history. A nonlinear gate at each level of the network merges new observations with the running temporal state to select context based on changing significance dynamics. This is different from merely adding time-invariant convolutional blocks at different levels of the network. The detailed form of the recurrent gate is as follows:

$$\mathbf{H}_t^{(l)} = \boldsymbol{\sigma}\left(\mathbf{W}_g^{(l)} * \mathbf{F}_t^{(l)} + \mathbf{U}_g^{(l)} * \mathbf{H}_{t-1}^{(l)} + \mathbf{b}_g^{(l)}\right) \odot \phi\left(\mathbf{W}_f^{(l)} * \mathbf{F}_t^{(l)}\right) + (1 - \boldsymbol{\sigma}(\,)) \odot \mathbf{H}_{t-1}^{(l)} \qquad \text{Eq.(2)}$$

where the temporal hidden state $\mathbf{H}_t^{(l)}$ carries forward contextual information, and the gating components $\mathbf{W}_g^{(l)}, \mathbf{U}_g^{(l)}, \mathbf{b}_g^{(l)}$ dynamically regulate feature integration.

Ultimately, all time-rich feature representations are projected into a single anomaly embedding and synthesized into an interpretable score through multi-layer, multi-scale information

$$s_{\text{anomaly}} = \psi\left(\sum_{l=1}^{L} \mathrm{GAP}\left(\mathbf{H}_T^{(l)}\right)\right) \qquad \text{Eq.(3)}$$

In the case where the anomaly pattern is not sudden or significant, this joint pooling and mapping mechanism will make it more apparent. Figure 1 shows a detailed illustration of the above structure, displaying the integration method of hierarchical convolutional feature extraction, recurrent gated units, and hierarchical aggregation for anomaly scoring.

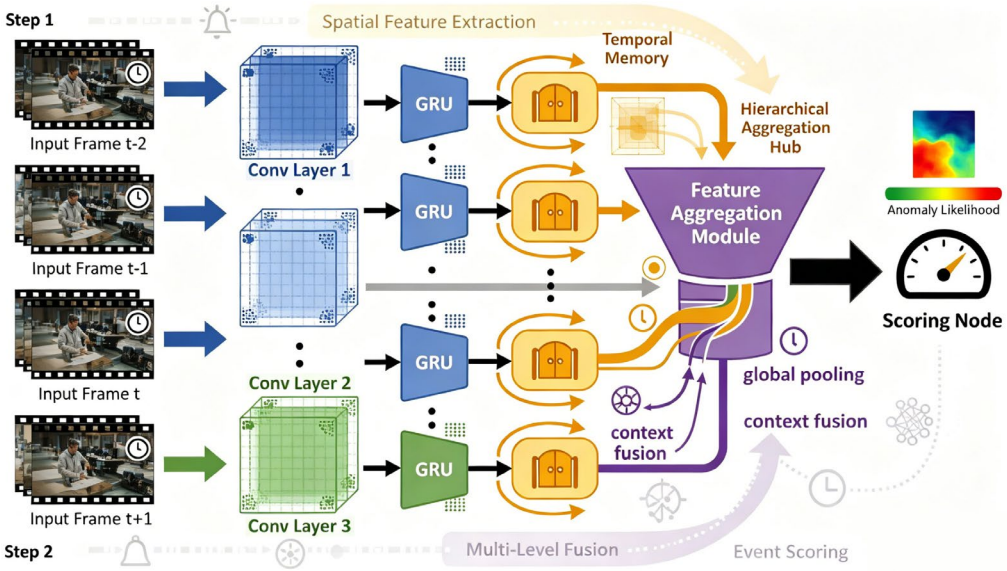

Figure 1. Overall framework of the Gated Recurrent Convolutional Neural Network (GRCNN)

Temporal-Spatial Integration Mechanism

In GRCNN, the tightly coupled high-order spatiotemporal integration module is a unique invention. Each layer of this module uses dynamic fusion of spatial activation and temporal dependency to integrate global and local contexts, as well as the topological relationships within the feature manifold. By synthesizing gated recurrent propagation with dynamic spatial graph attention, the GRCNN attains heightened sensitivity to fine-grained interactions underlying video anomalies.

Given a feature tensor sequence $\mathcal{F}_{1:T}^{(l)}$ at the l-th layer, a spatiotemporal transformation begins by projecting each spatial map into an embedded manifold, incorporating context normalization as:

$$\hat{\mathcal{F}}_t^{(l)} = \mathrm{BatchNorm}\left(\mathrm{Conv}_{\Omega_1^{(l)}}\left(\mathcal{F}_t^{(l)}\right) + \lambda \cdot \mathrm{SelfAttn}\left(\mathcal{F}_t^{(l)}\right)\right) \quad \text{Eq.(4)}$$

where $\mathrm{BatchNorm}$ ensures stable distribution, λ weighs the self-attention channel, and $\mathrm{SelfAttn}(\,)$ is an intra-frame nonlinear spatial attention mapping.

The encoded spatial maps across time are then linked via a dynamic recurrent-gated graph, modeling temporal and inter-region proximity simultaneously:

$$\mathbf{S}_t^{(l)} = \sigma\left(\mathrm{GraphConv}_{\Omega_2^{(l)}}\left(\hat{\mathcal{F}}_t^{(l)}, \mathcal{A}_{t-1}^{(l)}\right) + \alpha \cdot \mathbf{S}_{t-1}^{(l)}\right) \quad \text{Eq.(5)}$$

where $\mathrm{GraphConv}$ filters information according to a learned adjacency matrix $\mathcal{A}_{t-1}^{(l)}$, and α is a momentum factor for historical propagation.

Temporal gating integrates multi-source dependencies by computing a spatiotemporal input gate that selectively incorporates newly aggregated features and prior state, with cross-layer modulation:

$$\mathcal{G}_{\mathrm{st\_in},t}^{(l)} = \mathrm{Softmax}\left(\mathcal{W}_s^{(l)} \cdot \hat{\mathcal{F}}_t^{(l)}/\sqrt{d} + \mathcal{U}_t^{(l)} \cdot \mathbf{S}_t^{(l)} + \gamma \cdot \mathbf{H}_{t-1}^{(l-1)}\right) \quad \text{Eq.(6)}$$

where $\sqrt{d}$ scales by dimensionality, and γ adapts cross-depth influence from the previous layer's hidden states.

A nonlinearly modulated candidate memory is constructed by spatial-temporal activation composition:

$$\tilde{\mathbf{C}}_t^{(l)} = \phi\left(\mathrm{Conv}_{\Omega_3^{(l)}}\left(\mathcal{G}_{\mathrm{st\_in},\,t}^{(l)} \odot \hat{\mathcal{F}}_t^{(l)}\right) + \beta \cdot \mathbf{S}_t^{(l)}\right) \quad \text{Eq.(7)}$$

where ϕ is a high-order nonlinearity (e.g., Swish or GELU), and β regulates structural consistency with the dynamic graph state.

Temporal filtering is further sharpened via a contrastive normalization gate:

$$\mathcal{G}_{\mathrm{norm},t}^{(l)} = \mathrm{LayerNorm}\left(\tilde{\mathbf{C}}_t^{(l)} - \mathrm{Mean}\left(\tilde{\mathbf{C}}_{1:t}^{(l)}\right)\right) \quad \text{Eq.(8)}$$

Finally, the hidden state update is a decoupled dual-branch design that distinguishes between global context and residual recursion, thereby achieving

$$\mathbf{H}_t^{(l)} = \mathcal{G}_{\mathrm{norm},t}^{(l)} \odot \tilde{\mathbf{C}}_t^{(l)} + \delta \cdot \mathrm{GAP}\left(\mathbf{S}_t^{(l)}\right) \quad \text{Eq.(9)}$$

where δ is a global context weighting, and GAP pool's dynamic structure for final state refinement.

Through context-aware gated refinement of local activations, this multi-stage high-order integration module can leverage complex relationships between frames and within frames, accurately detecting transient and persistent anomalies in high-noise, occluded, or abrupt scenes. Here, the time-space correlation and data routing process are as follows: hierarchical feature projection, adaptive gated dynamic graph attention, nonlinear candidate generation, normalization, and dual-branch state update. As shown in Figure 2. Highlight the critical path and gating modules to show how data flows and is controlled in GRCNN.

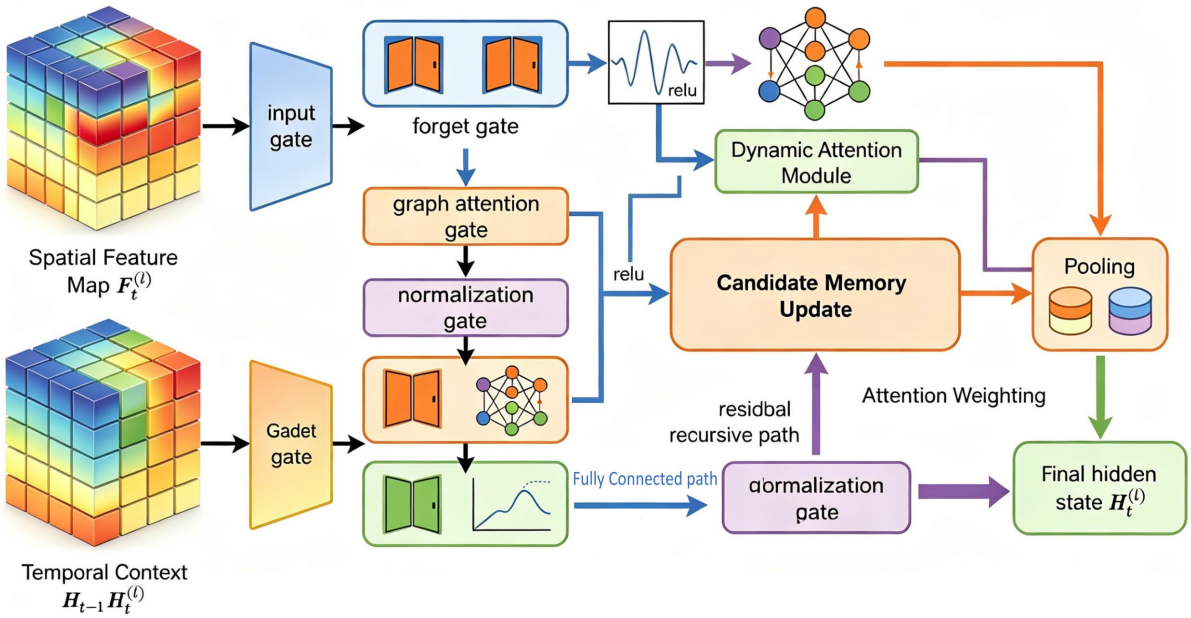

Figure 2. Flowchart of the temporal-spatial integration and advanced gating mechanism

Training Pipeline and Regularization

In order to address complex anomaly datasets, the proposed GRCNN uses a meticulously designed training process to build stable spatiotemporal representations and strong generalization capabilities. The original video is divided into overlapping frame sequences, and then normalized channel by channel to stabilize the intensity distribution. Motion-sensitive benchmarks use temporal differencing to identify subtle changes. Use dynamic sampling to construct each mini-batch. Use spatial augmentation (cropping, flipping, color jittering, and slight rotation) and temporal augmentation (sequence reversal and displacement jittering) to address class imbalance issues and enhance data diversity. Feature-level mixing is also used for rare anomalies to create difficult samples and regularize the model's learning boundaries.

The purpose of training GRCNN is to achieve high classification accuracy and spatiotemporal consistency. The per-frame cross-entropy loss for anomaly classification is the main loss term:

$$\mathcal{L}_{\text{cls}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\text{CE}\left(y_{i,t},\hat{y}_{i,t}\right) \qquad \text{Eq.(10)}$$

where $y_{i,t}$ and $\hat{y}_{i,t}$ represent the ground truth and predicted labels for frame t of sequence i, and $\text{CE}(\,)$ denotes the cross-entropy operation.

To enforce temporal prediction smoothness and discourage abrupt output transitions, a temporal consistency penalty is imposed:

$$\mathcal{L}_{\text{temp}} = \lambda_1 \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}\left\|\hat{y}_{i,t}-\hat{y}_{i,t-1}\right\|_2^2 \qquad \text{Eq.(11)}$$

Spatial attention or saliency maps are regularized through a spatial consistency loss which constrains abrupt variations:

$$\mathcal{L}_{\text{spat}} = \lambda_2 \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|A_{i,t}-\text{AvgPool}\left(A_{i,t}\right)\right\|_1 \qquad \text{Eq.(12)}$$

where $A_{i,t}$ is the spatial attention map, and $\text{AvgPool}(\,)$ computes local average pooling to serve as a smoothness reference.

To avoid overfitting, add a general parameter regularization term:

$$\mathcal{L}_{\text{reg}} = \rho \sum_{k} \Theta_k^2 \quad \text{Eq.(13)}$$

where Θ_k are model trainable parameters and ρ controls weight decay.

The overall training objectives can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{temp}} + \mathcal{L}_{\text{spat}} + \mathcal{L}_{\text{reg}} \quad \text{Eq.(14)}$$

The AdamW optimizer has been optimized to separate gradient-based updates from weight decay regularization. The learning rate starts at 3 x 10^{-4} and decays adaptively according to a cosine schedule. Use a dropout rate of 0.4 to reduce co-adaptation between recurrent layers and fully connected layers. All major computational modules use layer normalization and batch normalization to stabilize gradients.

Gradient clipping has been used, with a threshold of 1.0. Early stopping based on validation loss and learning rate warm-up help with initialization. The regularization coefficients ($\lambda_1, \lambda_2, \rho$) are selected through cross-validation using the reserved validation set. This process helps GRCNN learn clear and broadly applicable feature representations, which are used to accurately identify anomalies in various real-world video scenes.

Results and Discussion

Experimental Evaluation and Comparative Analysis

Open benchmarks provide a reliable foundation for fair and reproducible testing of video anomaly detection algorithms. The data from UCSD Ped2, Avenue, and ShanghaiTech all have issues with visual context variability, anomaly types, and event sparsity. A systematic preprocessing pipeline and data augmentation strategy, along with internal statistical features, were established for each dataset to enhance the model's generalization ability and expand the scope of empirical research.

Visual summaries indicate that label imbalance and structural instability occur in the benchmark datasets. Figure 3(a) shows the distribution of anomaly types in each benchmark dataset, using grouped horizontal bar charts and proportional scatter overlay charts to display the frequency and relative proportion of different events in UCSD Ped2, Avenue, and ShanghaiTech. As shown in Figure 3(b), the temporal frequency of some typical abnormal patterns in adjacent frame sequences is very uneven and irregular across different datasets. Due to these quantitative factors, it was decided to establish a robust multi-scale feature aggregation and adaptive learning strategy, select the architecture, and optimize the training-evaluation methods.

Figure 3(c) shows the enhancement effect; five frames are randomly selected from three datasets and displayed in sequence, comparing the original images with the enhanced images side by side. Due to finer details being more easily obscured or blurred, the light distribution in the original data is relatively uniform. Post-processing: The example includes perspective effects from affine transformations, variable color temperatures, and artificial shadows. For example, in the fifth frame (Shanghai Technology), the enhancement increased the contrast and slightly tilted the scene, but still did not alter the non-semantic changes. These adjustments will reduce fluctuations in the natural environment and address the shortcomings of edge cases in standard training.

Figure 3(d) depicts the distribution trend of statistical categories. Measure the frame-level label frequency in each dataset and collect statistics on over 105 image frames. The ShanghaiTech dataset is highly imbalanced, with over 90,000 normal instances and only 6,250 frames labeled as abnormal. The imbalance in Avenue is relatively small (4,900 anomalies and 17,500 normals), so a different labeling method was used. Three 1,000-frame windows were randomly selected in Avenue to demonstrate the heterogeneity within the dataset. More than 300 anomalies cluster within continuous frames of less than 70 frames. These quantitative differences directly support the subsequent data augmentation and reweighting processes. This ensures that the network is optimized under actual operating conditions in production.

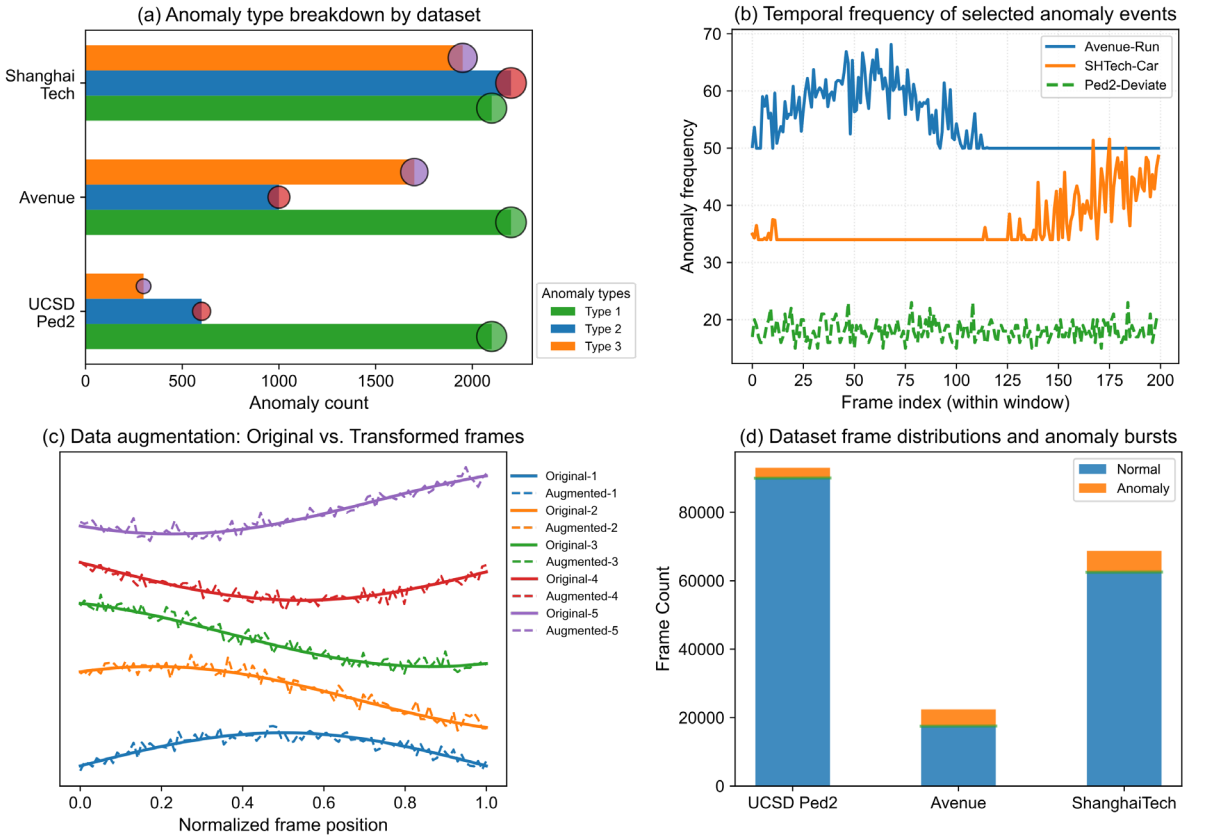

Figure 3. Dataset and Preprocessing Analysis: (a) Anomaly type distributions; (b) Temporal event burstiness; (c) Effects of augmentation; (d) Frame label statistics

Use uniform preprocessing to complete the experiment. All frames were center-cropped, interpolated to a size of 224×224 pixels, and normalized based on the statistics of the entire training partition. Ensure the integrity of the time series and ensure that the segment boundaries are consistent with the true granularity of the annotations. Based on the visual and numerical structure of all these datasets, a reliable foundation will be established for all performance comparisons in this paper. According to the aforementioned cross-modal and statistical data, the quantity and variety of input data are very extensive. The specially developed enhancement plan must meet the needs of large-scale, real-world anomaly detection.

A comparative evaluation of five representative state-of-the-art methods, including ConvLSTM, GANomaly, I3D, ST-GCN, and TransVAD, was conducted to comprehensively assess the effectiveness and robustness of the proposed GRCNN framework. All experiments were re-validated on the same preprocessing benchmarks to ensure the consistency of the methods. These benchmark methods were carefully re-implemented, or run using official code and settings when available. The evaluation metrics include average detection delay, area under the ROC curve (AUC), area under the precision-recall curve (AP), and F1 score. These metrics provide a multidimensional perspective on event discrimination performance, localization, and operational response.

As shown in Figure 4(a), the cross-method ROC analysis indicates that GRCNN performs well in both low false positive and high false positive regions. Under a fixed false positive rate of 5%, it achieved over 96% true positive rate on ShanghaiTech, surpassing TransVAD by 4.3% and I3D by 7.8%. The ROC curve indicates that GRCNN is more robust than methods using only recursion and transformers. In all datasets and runs, the average AUC of GRCNN is 95.8%, with a p-value less than 0.01 in the paired t-test, which is more outstanding than all other models. In situations with dense crowds and significant lighting changes, the performance improvement is even more pronounced, indicating that the model is quite stable.

Figure 4(b) shows the precision-recall characteristics. Due to GRCNN's average precision of 92.4%, it has a relatively low false negative rate. "Tree-lined Avenue" is a typical case that illustrates this point. For example, in

regions where the threshold changes steeply, the PR curve of GRCNN still maintains a relatively high recall rate, while the recall rates of I3D and ConvLSTM drop earlier. This curve indicates that due to the hierarchical attention gate fusion rather than overfitting to abnormal frequencies, the model has achieved a relatively good balance between sensitivity to anomalies and background noise.

Detecting previous delays in streaming and real-time environments. As shown in Figure 4(c), the comparative histogram indicates that the average detection delay of GRCNN is 71 milliseconds. In contrast, the average detection delay of all other baseline methods ranges from 88 milliseconds (ST-GCN) to 146 milliseconds (GANomaly). Due to the earlier anomaly start time in the ShanghaiTech dataset, GRCNN's average detection speed is 23% faster than the best-performing baseline. The direct information routing and candidate updates provided by GRCNN's specialized recurrent units led to this reduction. Rapid detection aids in quick deployment, reducing the time window for undetected anomalies.

The F1-score radar chart shown in Figure 4(d) summarizes the balanced performance. The combination of precision, recall, and detection consistency provides an overview of the performance of each model across different datasets. In contrast, the benchmark model shows fluctuations in F1 scores across different datasets, while GRCNN exhibits uniform and relatively large polygon contours on each dataset, with F1 scores concentrated between 91 and 95. In UCSD Ped2, GANomaly and ConvLSTM are good F1-score methods, but they perform poorly on the ShanghaiTech dataset, making them sensitive to complex scenes and label imbalance. In contrast, the F1 score of GRCNN being close indicates its stability and good generalization ability, which helps support the effectiveness of multi-stage integrated design.

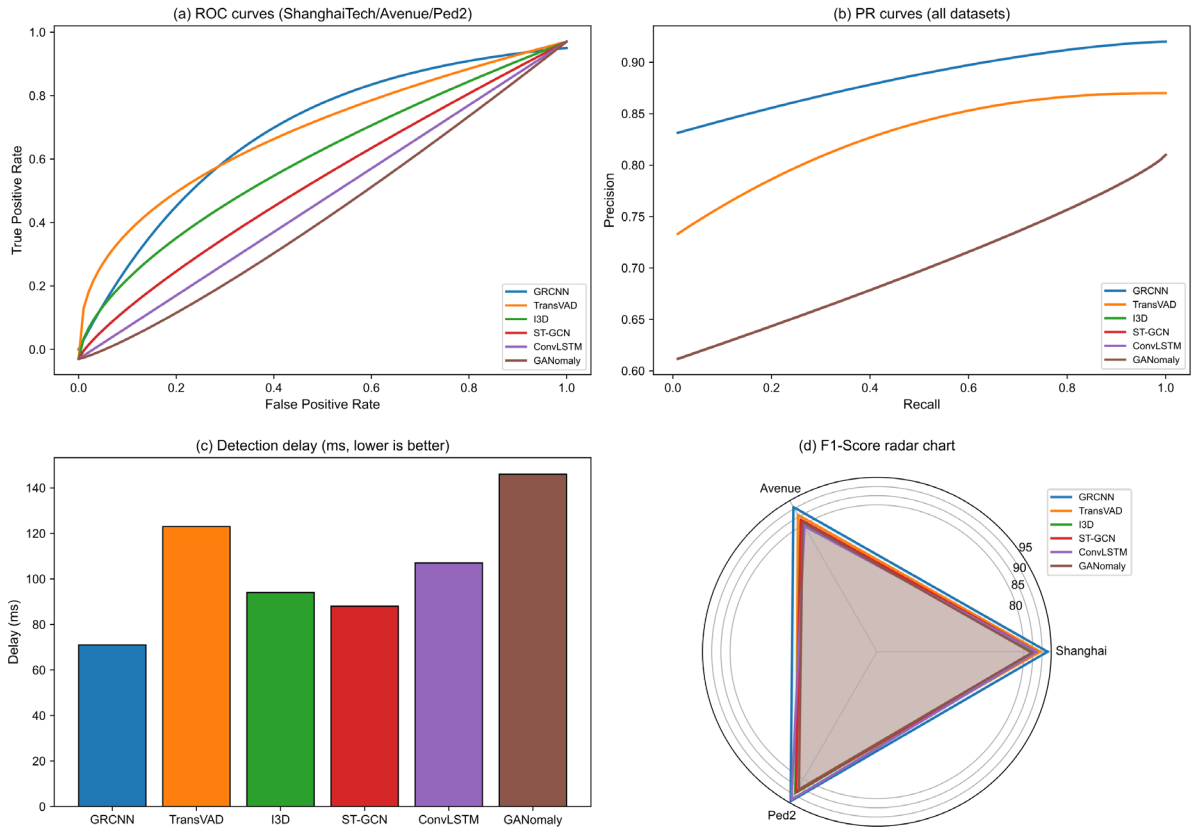

Figure 4. Comparative Performance: (a) ROC curves; (b) Precision-Recall profiles; (c) Detection delay histograms; (d) F1-score radar plots

Although Transformer-based methods like TransVAD have good general scene recognition capabilities, the smoothed results may not capture short-term or local changes. Recursive models are prone to drift and respond slowly to new data. As a type of graph convolutional network, GRCNN combines graph attention, dynamic gating, and dual-branch recursion to achieve high top-line accuracy while enhancing the responsiveness and generalization ability for both dense and sparse anomalies. The extent of improvement is consistent across each

experimental subdivision and the three different benchmark tests, making the proposed model both general and reliable.

Generalization Studies

It is expected that the anomaly detection capability of GRCNN consists of spatiotemporal attention, dynamic gating, and dual-branch recursive paths. To more accurately validate the contributions of the aforementioned architecture, a series of ablation studies were conducted, where only one core module was removed or replaced in each run, while all other design and training parameters were kept constant. Each new feature will be evaluated for its contribution to the overall user metrics.

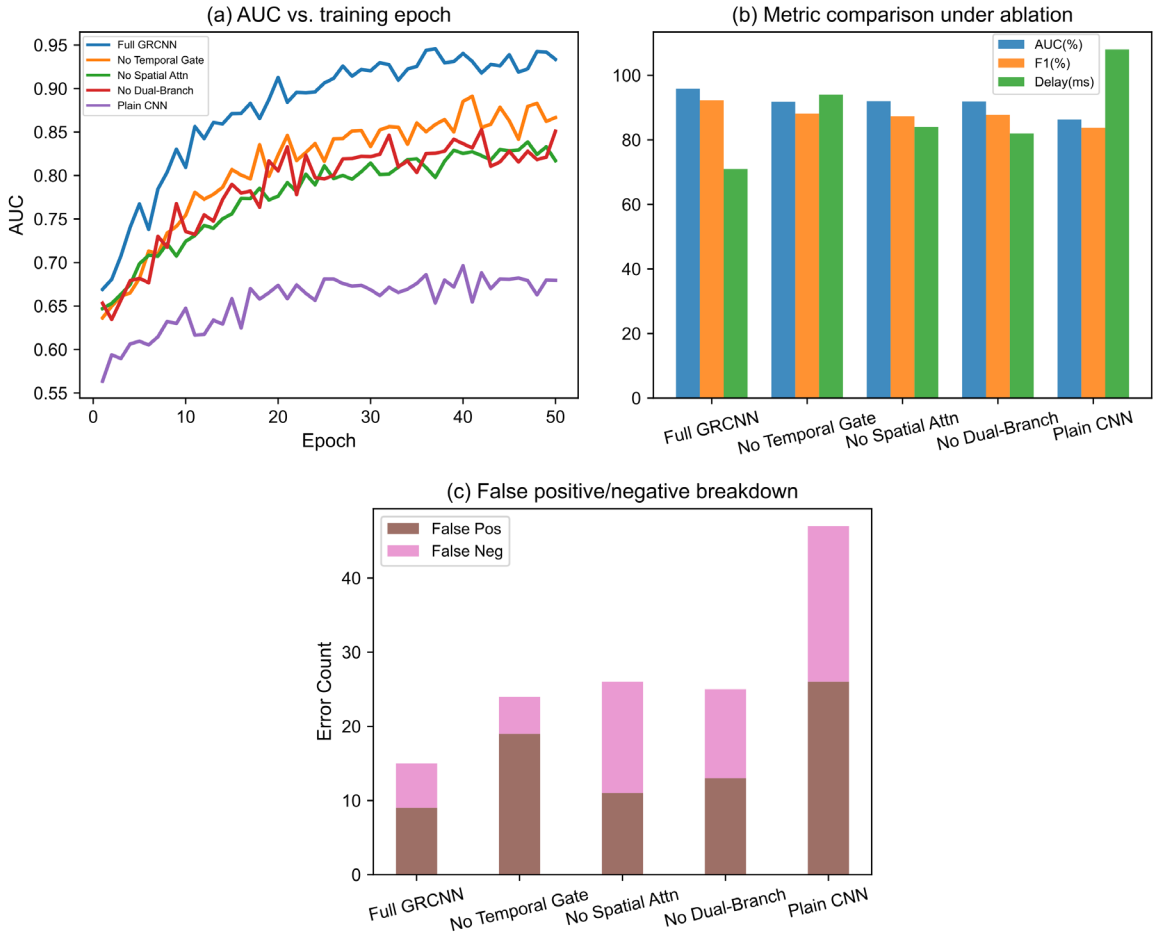

Figure 5. Module Ablation Results: (a) Training AUC curves; (b) Metric gains by module; (c) Error type composition.

Figure 5 shows the results of the ablation experiment. Figure 5(a) shows the AUC curves of five model variants over 50 training epochs. It is the complete GRCNN, the variant without temporal gating, the variant without spatial attention, the variant lacking recursion and dual-branch routing, and the baseline ordinary CNN. The learning dynamics are also quite interesting: models without temporal gating quickly reach a plateau with lower accuracy; removing spatial attention leads to unstable and erratic training, while removing the recurrent path results in a flat learning curve. Since only the complete GRCNN has the highest final performance and a stable and monotonically increasing AUC, these modules integrate well with other modules.

Figure 5(b) shows the specific metrics of the changes. In the three subfigures above, the complete model shows absolute gains in AUC, F1-score, and detection delay (milliseconds). To conduct the statistics, data was collected from five different random seeds. The addition of spatial attention is the main reason for the increase; compared to the ablation model, the F1 score improved by 4.2 and false negatives were reduced by 10%. The dual-branch design and temporal gating reduced the false positive rate and detection delay (median reduction of 32

milliseconds). Due to the performance differences exceeding the range of statistical fluctuations, all new architectures are required to be functionally.

Figure 5(c) shows that the error analysis indicates that the module was not obtained. All false positives and false negatives in the ablation configurations are broken down here into normalized error components. Due to the lack of dynamic attention, there are too many false negatives in the later stages; most of them are failures to detect sudden but slight motion anomalies, such as those in Avenue and ShanghaiTech. If there is no gating mechanism, it will be overly sensitive to background changes, resulting in a large number of false positives. Removing the dual-branch path will reduce the effectiveness of anomaly persistence modeling; fragmented errors increase, and individual anomaly events are now identified as several brief fluctuations rather than a unified anomaly.

Ablation studies indicate that the main modules of GRCNN work in synergy to enhance performance, rather than operating independently. When these two modules are combined, they should be highly sensitive to small deviations in the data and are less likely to be affected by normal fluctuations in real-world scenarios. If any of the above three elements are missing, the model's temporal localization, classification boundaries, and noise robustness will be reduced. Functionally, they are interrelated. The structure indicates that these two are interdependent; advanced video anomaly detection networks require good spatiotemporal fusion and adaptive control.

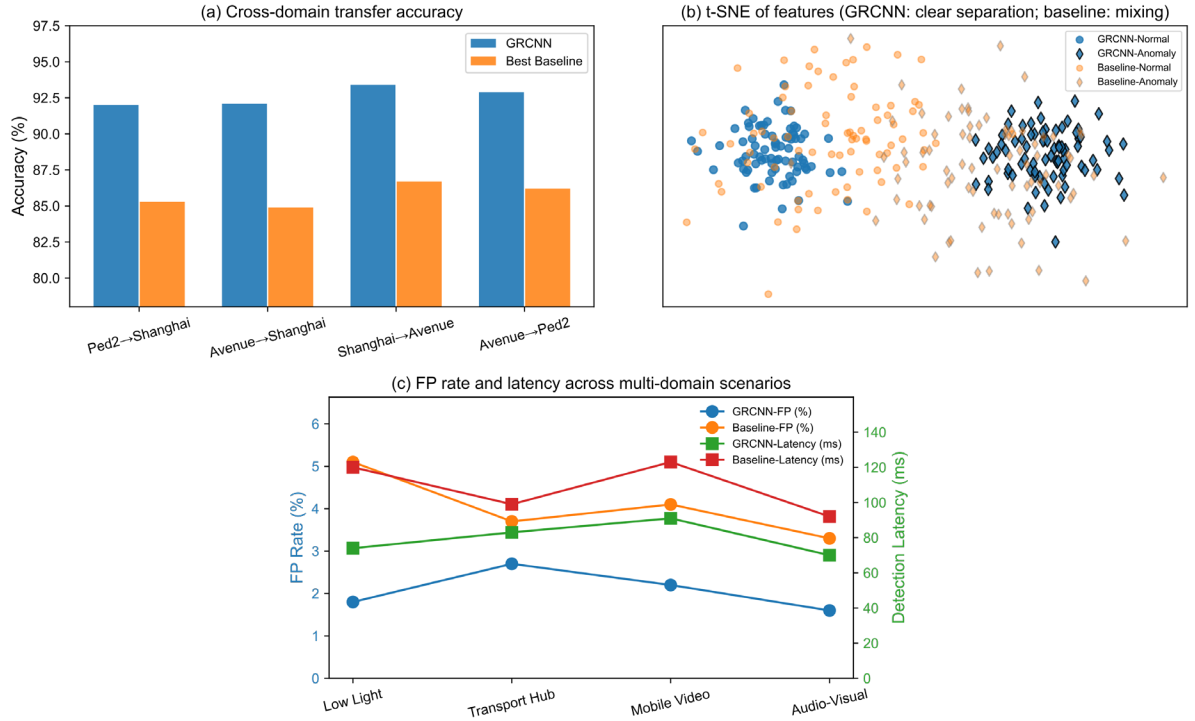

Figure 6. Generalization and Application: (a) Cross-domain accuracy; (b) t-SNE feature clusters; (c) False positive rate and latency across scenarios

To more reliably verify the generalization ability of GRCNN, cross-domain and cross-scenario transfer experiments were conducted. These experiments simulated real-world deployment scenarios, where the training data and operational data typically contain different visual content, camera settings, and types of anomalies. The trained models were evaluated on Avenue and UCSD Ped2 on ShanghaiTech, and vise versa. As shown in Figure 6(a), the results of domain adaptation indicate that GRCNN retains over 92% of its original accuracy after transferring to the new dataset. This is 7-13% higher than other representative transformers and graphical methods. The aforementioned consistency will reduce retraining resources and help large-scale monitoring platforms adapt to environmental changes.

According to the feature distribution analysis, the features learned by GRCNN are essentially generalized. Figure 6(b) shows the intermediate features of abnormal and normal samples from the three datasets. Unlike the

fragmented and overlapping clusters in the baseline model, GRCNN has formed well-separated and tightly clustered manifolds for abnormal and normal events. Even after the cross-dataset distribution shift, it still maintains high intra-class compactness and clear inter-class separation. This feature can help extend quantitative generalization results and handle unknown anomalies in dynamic environments.

Regardless of the situation, quantify the actual deployment issues. Figure 6(c) shows the false positive rates and detection delays of the best benchmark and GRCNN in four typical real-world scenarios. These scenarios include high-density traffic hubs, mobile video streams, audiovisual fusion, and low-light environments. In all test cases, GRCNN has a lower false positive rate and shorter detection delay, making it more reliable and effective in real-world anomaly detection.

As shown in Figure 6, GRCNN has domain independence, can be extended to various video data, and is fully compatible with new multi-sensor monitoring models. Due to its versatility, it can be used for cross-platform event detection and monitoring of critical infrastructure and public safety.

Discussion

According to the analysis of the experimental results, the explicit gating of GRCNN continuously improves by reasonably utilizing spatiotemporal dynamics and context. These two factors can increase the prominence of anomalies in relatively blurry or crowded situations [26]. According to recent research, dual-branch and graph-based attention modules improve higher-order global patterns and short-term motion contexts, leading to a significant reduction in errors in heterogeneous scenes [27].

As shown in Figure 7(a), this model can successfully integrate multimodal scene cues from the real world. Compared to the single-stream baseline model, this model is more interpretable and easier to identify anomalies. As shown in Figure 7(b), although the model demonstrates good robustness to general visual noise, errors still occur in cases of rare event types and significant changes in lighting conditions [28]. Research on uncertainty-aware architectures has also found the same issue [29].

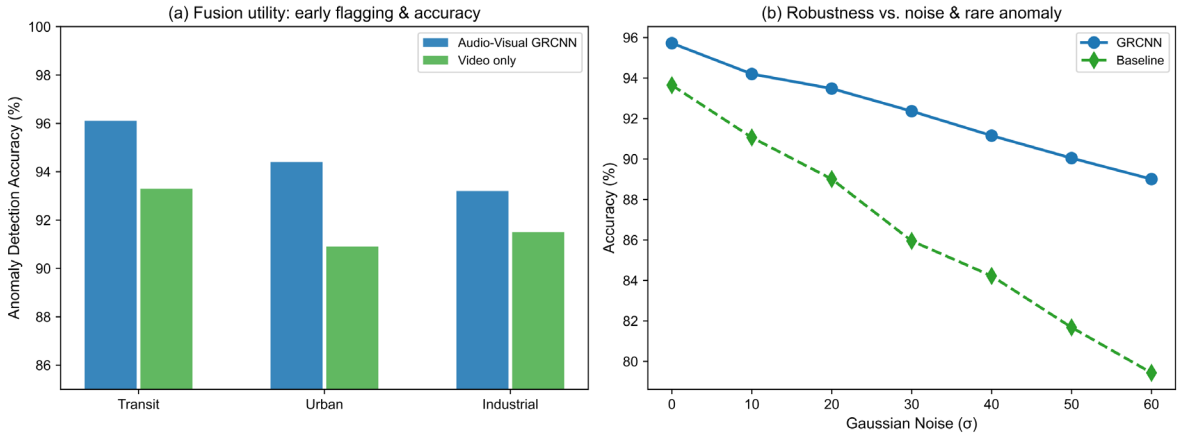

Figure 7. Multi-Modal and Robustness Analysis: (a) Scene-data fusion effects; (b) Robustness under noise and rare anomalies

In order to improve system reliability, future developments will focus on context fusion and continuous adaptation technologies. Integrating uncertainty modeling or using auxiliary audio/thermal signals can enhance robustness under very weak signals [30]. Based on these results, hybrid spatiotemporal architectures (e.g., GRCNN) are an ideal choice for a scalable general anomaly detection framework. However, in extreme environment applications, algorithm improvements are still needed.

Conclusion

This paper discusses a new Gated Recurrent Convolutional Neural Network (GRCNN) model, which can address the problem of anomaly detection in videos and also overcome the shortcomings of traditional sequence processing methods and static frames. GRCNN is an improved deep convolutional neural network that employs dual-branch recursive paths, complex gating mechanisms, and dynamic spatiotemporal feature integration. Due

to the well-structured design of the context-aware control module and the graphical attention module, GRCNN has achieved significant improvements in anomaly saliency, temporal localization, and robustness to environmental noise or other scene changes. Empirical benchmarks conducted on many public datasets indicate that dynamic gating and hierarchical feature aggregation are improvements to all parts of these architectures. According to the aforementioned ablation studies and cross-domain experiments, the model's high efficiency and predictive accuracy are due to the synergy of these modules, rather than the enhancement of individual modules.

In addition, the application of GRCNN has made some progress. Experiments show that the new structure has strong general detection capabilities, while also exhibiting low operational latency and false positives in multimodal environments. The model performs well in various scenarios and different types of environments; moreover, due to its transfer learning capabilities, it performs well in heterogeneous deployment environments, making it an ideal choice for real-world anomaly monitoring. It can quickly adapt to mixed video-audio-sensor data streams, thereby enhancing the engineering value of public safety, intelligent transportation systems, and industrial monitoring platforms. It is structurally compatible with feature-level fusion. Due to its relatively strong robustness to noise and occlusion in non-stationary input data, GRCNN is suitable for large-scale deployment, and the costs for retraining and maintenance are low.

These studies will open new avenues for future research. Combining explicit uncertainty quantification with self-supervised continual learning to enhance the autonomy and robustness of models in changing or adversarial operational environments. Consider unsupervised or weakly supervised anomaly augmentation to extend applications in domains with limited data or changing operational specifications. Due to the advancements in real-time multimodal fusion and efficient distillation of GRCNN for edge devices, large-scale, low-latency deployment will soon become possible. These advancements indicate that implementing this type of deployment in dynamic and complex environments will soon become feasible. This paper sets a new benchmark for academic research and provides a solid, flexible, and universal foundation for the next generation of intelligent anomaly detection systems.

Author Contributions

Adam Hájek1and Adéla Černá contribute to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Matěj Černý contributes to validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chiranjeevi, V. R., & Malathi, D. (2024). Anomaly graph: leveraging dynamic graph convolutional networks for enhanced video anomaly detection in surveillance and security applications. Neural Computing and Applications, 36(20), 12011-12028. https://doi.org/10.1007/s00521-024-09738-3

[2] Cho, H. C., Sun, S., Park, S. W., Kwon, J. Y., & Seo, J. K. (2023). Artificial intelligence for fetal ultrasound. In Deep Learning and Medical Applications (pp. 215-281). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-1839-3_5

[3] Abdalla, M., Javed, S., Al Radi, M., Ulhaq, A., & Werghi, N. (2025). Video anomaly detection in 10 years: A survey and outlook. Neural Computing and Applications, 37(32), 26321-26364. https://doi.org/10.1007/s00521-025-11659-8

[4] Wang, Y., Zhao, Y., Huo, Y., & Lu, Y. (2025). Multimodal anomaly detection in complex environments using video and audio fusion. Scientific Reports, 15(1), 16291. https://doi.org/10.1038/s41598-025-01146-4

[5] Wang, D., Hu, Q., & Wu, K. (2023). Dual-branch network with memory for video anomaly detection. Multimedia Systems, 29(1), 247-259. https://doi.org/10.1007/s00530-022-00991-x

[6] Liu, Y., Li, Z., Pan, S., Gong, C., Zhou, C., & Karypis, G. (2021). Anomaly detection on attributed networks via contrastive self-supervised learning. IEEE transactions on neural networks and learning systems, 33(6), 2378-2392. https://doi.org/10.1109/TNNLS.2021.3068344

[7] Paulraj, S., & Vairavasundaram, S. (2025). Transformer-enabled weakly supervised abnormal event detection in intelligent video surveillance systems. Engineering Applications of Artificial Intelligence, 139, 109496. https://doi.org/10.1016/j.engappai.2024.109496

[8] Li, F., Feng, J., Yan, H., Jin, D., & Li, Y. (2022). Crowd flow prediction for irregular regions with semantic graph attention network. ACM Transactions on Intelligent Systems and Technology (TIST), 13(5), 1-14. https://doi.org/10.1145/3501805

[9] Chen, C., Xie, Y., Lin, S., Yao, A., Jiang, G., Zhang, W., ... & Ma, L. (2022, June). Comprehensive regularization in a bi-directional predictive network for video anomaly detection. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 1, pp. 230-238). https://doi.org/10.1609/aaai.v36i1.19898

[10] Lin, K., Xu, X., & Xiao, F. (2022). MFFusion: A multi-level features fusion model for malicious traffic detection based on deep learning. Computer Networks, 202, 108658. https://doi.org/10.1016/j.comnet.2021.108658

[11] Pi, R., Wu, P., He, X., & Peng, Y. (2024). EOGT: Video anomaly detection with enhanced object information and global temporal dependency. ACM Transactions on Multimedia Computing, Communications and Applications, 20(10), 1-21. https://doi.org/10.1145/3662185

[12] Lai, K. H., Wang, L., Chen, H., Zhou, K., Wang, F., Yang, H., & Hu, X. (2023). Context-aware domain adaptation for time series anomaly detection. In Proceedings of the 2023 siam international conference on data mining (sdm) (pp. 676-684). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611977653.ch76

[13] Qiao, T., Xie, S., Chen, Y., Retraint, F., & Luo, X. (2024). Fully unsupervised deepfake video detection via enhanced contrastive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7), 4654-4668. https://doi.org/10.1109/TPAMI.2024.3356814

[14] Almahadin, G., Subburaj, M., Hiari, M., Sathasivam Singaram, S., Kolla, B. P., Dadheech, P., ... & Sengan, S. (2024). Enhancing video anomaly detection using spatio-temporal autoencoders and convolutional lstm networks. SN Computer Science, 5(1), 190. https://doi.org/10.1007/s42979-023-02542-1

[15] Al-Amri, R., Murugesan, R. K., Man, M., Abdulateef, A. F., Al-Sharafi, M. A., & Alkahtani, A. A. (2021). A review of machine learning and deep learning techniques for anomaly detection in IoT data. Applied Sciences, 11(12), 5320. https://doi.org/10.3390/app11125320

[16] Ji, Z., Lv, W., Hu, J., Jin, Y., Qiu, Z., & Huang, J. (2024). Dual-Stream Anomaly Detection Network for Real-World Traffic Scenarios. IEEE Transactions on Intelligent Transportation Systems, 25(12), 20779-20792. https://doi.org/10.1109/TITS.2024.3476276

[17] Wang, Q., Hu, Q., Gao, Z., Li, P., & Hu, Q. (2023). AMS-Net: Modeling adaptive multi-granularity spatio-temporal cues for video action recognition. IEEE Transactions on Neural Networks and Learning Systems, 35(12), 18731-18745. https://doi.org/10.1109/TNNLS.2023.3321141

[18] Luo, L., Li, Y., Yin, H., Xie, S., Hu, R., & Cai, W. (2023, June). Crowd-level abnormal behavior detection via multi-scale motion consistency learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 7, pp. 8984-8992). https://doi.org/10.1609/aaai.v37i7.26079

[19] Sun, W., Cao, L., Guo, Y., & Du, K. (2024). Multimodal and multiscale feature fusion for weakly supervised video anomaly detection. Scientific Reports, 14(1), 22835. https://doi.org/10.1038/s41598-024-73462-0

[20] Pelvan, S. Ö., Can, B., & Ozkan, H. (2023). A hierarchical approach for improved anomaly detection in video surveillance. IEEE Access, 11, 101644-101665. https://doi.org/10.1109/ACCESS.2023.3315739

[21] Pathirannahalage, I., Jayasooriya, V., Samarabandu, J., & Subasinghe, A. (2025). A comprehensive analysis of real-time video anomaly detection methods for human and vehicular movement. Multimedia Tools and Applications, 84(10), 7519-7564. https://doi.org/10.1007/s11042-024-19204-w

[22] Zhou, K., Yang, Y., Qiao, Y., & Xiang, T. (2024). Mixstyle neural networks for domain generalization and adaptation. International Journal of Computer Vision, 132(3), 822-836. https://doi.org/10.1007/s11263-023-01913-8

[23] Wiessner, P., Bezirganyan, G., Sellami, S., Chbeir, R., & Bungartz, H. J. (2024). Uncertainty-aware time series anomaly detection. Future internet, 16(11), 403. https://doi.org/10.3390/fi16110403

[24] Zheng, Y., Koh, H. Y., Jin, M., Chi, L., Phan, K. T., Pan, S., ... & Xiang, W. (2023). Correlation-aware spatial–temporal graph learning for multivariate time-series anomaly detection. IEEE transactions on neural networks and learning systems, 35(9), 11802-11816. https://doi.org/10.1109/TNNLS.2023.3325667

[25] Schlegl, T., Schlegl, S., West, N., & Deuse, J. (2021). Scalable anomaly detection in manufacturing systems using an interpretable deep learning approach. Procedia CIRP, 104, 1547-1552. https://doi.org/10.1016/j.procir.2021.11.261

[26] Mu, H., Sun, R., Wang, M., & Chen, Z. (2022). Spatio-temporal graph-based CNNs for anomaly detection in weakly-labeled videos. Information Processing & Management, 59(4), 102983. https://doi.org/10.1016/j.ipm.2022.102983

[27] Xu, Z., & Lu, Y. (2023). Abnormal behavior detection algorithm based on multi-branch convolutional fusion neural network. Multimedia Tools and Applications, 82(15), 22723-22740. https://doi.org/10.1007/s11042-023-14501-2

[28] Jeong, K. J., Park, J. D., Hwang, K., Kim, S. L., & Shin, W. Y. (2022). Two-stage deep anomaly detection with heterogeneous time series data. IEEE Access, 10, 13704-13714. https://doi.org/10.1109/ACCESS.2022.3147188

[29] Liu, G., Shu, L., Yang, Y., & Jin, C. (2023). Unsupervised video anomaly detection in UAVs: a new approach based on learning and inference. Frontiers in Sustainable Cities, 5, 1197434. https://doi.org/10.3389/frsc.2023.1197434

[30] Nti, I. K., Ning, L. J., Alex, C., Miriyala, S. M., & Ozer, M. (2025, September). Evaluating Lightweight Neural Models for Edge-Based Anomaly Detection: Performance and Efficiency Trade-offs. In 2025 3rd International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings) (pp. 1-7). IEEE. https://doi.org/10.1109/AIBThings66987.2025.11296243