Senthil *et al.*
EURASIP Journal on Image and Video Processing (2025) 2025:21
https://doi.org/10.1186/s13640-025-00685-0

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Abnormality-Aware Fused Attention Model with Global Density Joined Network for unusual activity detection in surveillance video

D. Siva Senthil[1*], R. Jagadish Vijay[2], A. Aalan Babu[3] and Shachi Mall[4]

*Correspondence:
sivasenthildss@gmail.com

[1] Department of Computer Science and Engineering, Arunachala College of Engineering for Women, Vellichanthai, Tamilnadu 629203, India
[2] Department of Electrical and Electronics Engineering, Amrita College of Engineering and Technology, Nagercoil, Tamilnadu 629901, India
[3] Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Tiruchirappalli Campus, Tiruchirappalli, Tamilnadu 621105, India
[4] Department of Computer Science and Engineering, Galgotias University, Greater Noida, India

Abstract

Detecting abnormal events in videos is essential for effective anomaly detection in surveillance environments. In this work, we propose a framework that combines the Abnormality-Aware Fused Attention Model (AAFAM) with the Global Density Joined Network (GDJNet) to enhance abnormal event detection. AAFAM uses spatial and channel-wise attention mechanisms to focus on anomalous regions in feature maps. It suppresses irrelevant background information. GDJNet, on the other hand, captures both local and global spatial relationships through density estimation. It allows the model to learn object distributions and co-occurrence patterns at multiple scales. To strengthen the model's performance, we integrate AAFAM and GDJNet using a fusion strategy that combines attention and density maps, resulting in highly discriminative feature representations. Experimental results show that the proposed AAFAM–GDJNet framework outperforms existing methods and achieves state-of-the-art performance.

Keywords: Abnormal event detection, AAFAM, GDJNet, Deep learning, Video analysis

1 Introduction

The analysis of human behavior has recently drawn significant attention in the fields of artificial intelligence and deep learning. In particular, automatic video anomaly detection has emerged as a key area of research [1–5]. The ability of a system to automatically identify deviations from normal behavior plays a crucial role in ensuring public safety and security [6–8]. The foremost aim of detecting abnormal events is to identify normal and abnormal patterns in the visual data. Whereas conventional methods of video analysis are primarily concerned with tracking or recognizing certain objects and behaviors, AED Abnormal Event Detection (AED), on the other hand, is about identifying those occurrences that are rare, unpredicted, and even dangerous [9–12].

Machine learning algorithms have been ranked among the most effective of the methods that have been widely used in AED under changing conditions [13–15]. These are mainly supervised methods like support vector machines and random forests, as well as unsupervised methods, such as clustering and anomaly detection. They do not perform directly on the raw data but instead run on images, videos, and time series [16].

In fact, deep learning models have been able to identify and localize abnormal events [17–20] better than any other method so far and have made the connection networks and the large datasets the basis for the discovery of complex patterns. This has resulted in a major breakthrough for their performance level to the point that they are considered state-of-the-art and are most often employed in real-world issues [16].

Abnormal event detection (AED) has evolved significantly over time; still, there are some roadblocks that hinder its transition to the next level. One of the biggest problems of high intra-class variability has been mentioned most frequently—abnormal situations change enormously depending on the surroundings and, therefore, it is extremely difficult for models to generalize. On top of that, the spatio-temporal complexity of the video data is very intriguing to researchers—it is difficult at the same time to extract spatial and temporal characteristics correctly. Until now, researchers have made important advances; however, they are still confronted with some difficulties. High intra-class variability is one of the main problems abnormally—events can vary a lot depending on the context; thus, making it difficult for models to generalize. Besides this, there are some challenges associated with the video-document spatio-temporal nature, which make the capturing of both spatial and temporal dependences a very complicated task.

To address these issues, we propose the Abnormality-Aware Fused Attention Model with Global Density Joined Network (AAFAM–GDJNet). This framework fuses local feature extraction with global context. The AAFAM module uses attention to focus on abnormal regions, by addressing the effects of intra-class variability. The GDJNet captures spatial density relationships to identify normal from abnormal motion patterns. The contributions of the proposed model are:

1. The AAFAM finds abnormal regions by highlighting important spatial features.
2. The GDJNet captures spatial relationships and object interactions at multiple scales, improving detection of anomalies across different environments.
3. The proposed model integrates AAFAM and GDJNet by fusing local feature enhancement (AAFAM) with global density estimation (GDJNet).
4. Extensive experiments on standard surveillance datasets demonstrate that AAFAM–GDJNet outperforms existing methods.

This article is organized as follows. Section 2 reviews methods for detecting abnormal behavior. Section 3 explains our proposed approach in detail. Section 4 shows the experimental results and demonstrates how well the method works. Section 5 provides the conclusion.

2 Related works

Recent advances in deep learning methodologies have shown substantial promise in video anomaly detection. Particularly, convolutional neural networks (CNNs) [21, 22] and recurrent neural networks (RNNs) [23] play a vital role due to their capacity to extract high-level semantic features. RNNs, especially LSTM and GRU models, are highly effective in capturing temporal patterns. Tan et al., introduced an improved GRU-based model that integrates the Random Forest algorithm with a fully connected layer network. This integration strictly controls the input of irrelevant features and

improves the model's fitting performance [24]. Recent studies have used two modified models based on Quantum Convolutional Neural Networks (Q-CNNs). These models combine quantum computing with traditional CNNs. They have been applied to classify real-time violent robberies and armed thefts [25]. Integrating Q-CNNs enables the model to tackle complex machine learning problems. However, Q-CNNs are still in the early stages of development and face practical challenges. Onyema et al. introduced the Slow-Fast Convolutional Neural Network (SF-CNN) framework, which adjusts its learning process according to the frame rate. It applies slow learning at lower frame rates to capture detailed spatial information. Fast learning is used at higher frame rates to analyze rapid temporal dynamics [26].

Another model involves autoencoders (AEs) that learn to reconstruct normal video patterns and flag discrepancies as anomalies. Reconstruction-based models, like Chong et al. [27] and Ionescu et al. [28], hold convolutional spatio-temporal autoencoders to encode appearance and motion features. However, these models often produce blurry reconstructions and fail to capture fine-grained details of fast-moving objects. Asad et al. introduced a dual-stream framework which integrates a 3D Convolutional Autoencoder (3D-CAE) to extract spatio-temporal features. They employed a One-Class Support Vector Machine (OCSVM) classifier to distinguish between normal and abnormal events [29]. Thakare et al. utilized 3D-CAE to extract spatio-temporal features from pseudo-labeled video sequences. This framework showed a promising improvement in UCSD Pedestrian, Shanghai Tech, and Avenue datasets [30]. Recent studies have improved human activity recognition through different learning frameworks. For example, a federated contrastive learning approach with feature-based distillation has been proposed to enhance generalization across distributed datasets [31]. An improved CNN architecture has been developed for accurate human detection and tracking in unconstrained environments [32]. Likewise, weakly supervised spatio-temporal models have demonstrated effective action localization in complex real-world scenarios [33]. Additionally, a generative framework based on the language of actions has been used to predict future human motion patterns from sensor data [34].

Attention mechanisms have been introduced to focus on salient spatial or temporal regions. Self-attention and temporal attention, as in Ada-Net [35], improve the representation of important cues across frames. As explained in reference [36], self-attention strikes an amazing balance between its ability to capture broad, long-range correlations within data and its effectiveness in both computational and statistical features. Zhang et al. [36] developed the idea of SAGAN, which cleverly integrates self-attention and Generative Adversarial Networks (GAN). By incorporating self-attention into the GAN framework, the generator gains the extraordinary capacity to generate images in which detailed details in one portion of the image seamlessly coincide with fine parts in distant regions. Li et al. [37] developed an approach that predicts subsequent frames in a video sequence and compares them to the actual frames. Significant discrepancies between the predicted and actual frames may indicate anomalies. An attention-based module is incorporated to improve the localization of these anomalies, and a memory addressing module is introduced to enhance predictive accuracy. In spite of the advancement, these models can suffer in highly cluttered scenes. To address these difficulties, the proposed

AAFAM–GDJNet introduces a hybrid architecture that combines the strengths of attention mechanisms and density-based modeling.

3 Methodology

3.1 The proposed Abnormality-Aware Fused Attention Model with Global Density Joined Network (AAFAM–GDJNet)

The proposed framework AAFAM–GDJNet enhances AED by integrating local feature development (AAFAM) along with global density estimation (GDJNet). This helps the model capture detailed spatial features while maintaining global context, improving accuracy and reducing false positives. Figure 1 depicts the proposed architecture. AAFAM refines features using spatial and channel attention to focus on anomalies. The GDJNet captures temporal dependencies and density variations. It clearly separates unusual events from regular background motion.

3.1.1 Architecture of the proposed AAFAM–GDJNet

The model consists of two main components:

AAFAM—extracts spatial features and improves them using spatial and channel attention mechanisms.

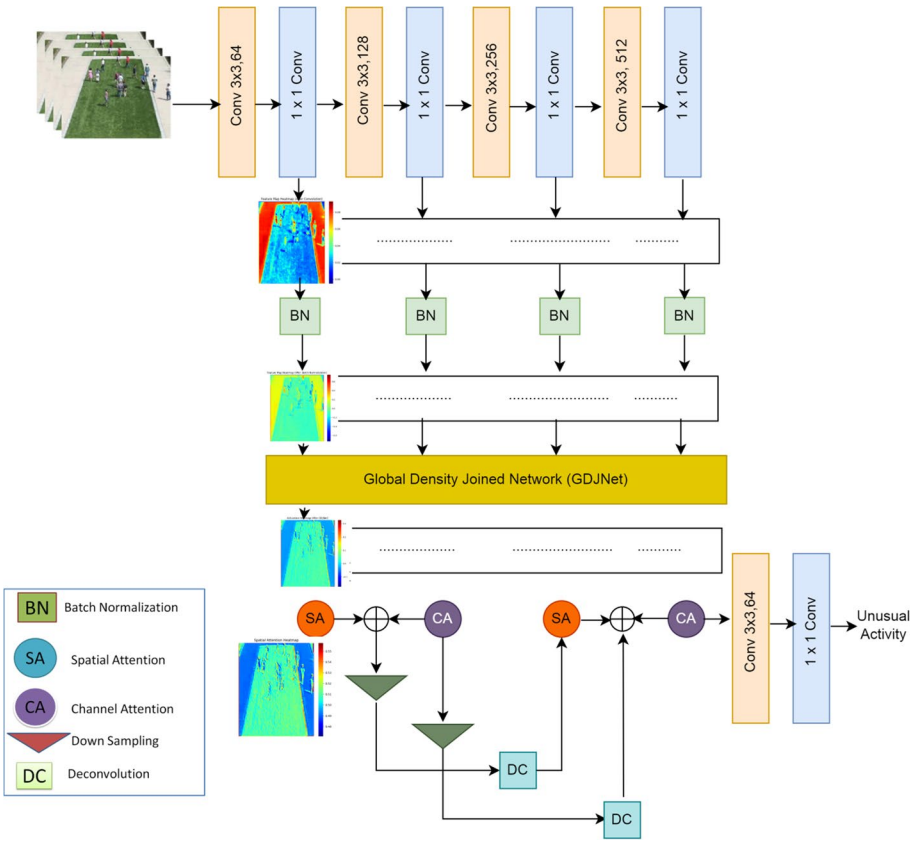

Fig. 1 Detailed architecture of the proposed structure

GDJNet—it learns local spatial relationships and global scene patterns through density estimation.

3.1.1.1 The proposed AAFAM The AAFAM typically employs convolutional layers to extract spatial features, while GDJNet may have its own feature extraction. Figure 2 depicts the GDJNet architecture. In AAFAM, convolutional layers are used as the model's first step in order to extract spatial characteristics from specific video frames. Local visual patterns are captured by these convolutional blocks, which then encode them as high-level feature representations. The model uses a fusion process over successive frames to include temporal information and capture motion patterns.

3.1.1.2 GDJNet: global density estimation and contextual modeling GDJNet uses small receptive field convolutions to capture fine details and local context. These local features provide fine-grained representations of regions of interest. GDJNet uses global context modeling to capture the entire scene information. This is done by using global average pooling (GAP). GDJNet ranks local feature representation along with global context modeling. After that, the resulting feature map (Fi) is converted into a one-dimensional vector by using bilinear interpolation. Next, the feature map is divided into four equal parts, as R1, R2, R3, and R4. Each part denotes a portion of the global information. The

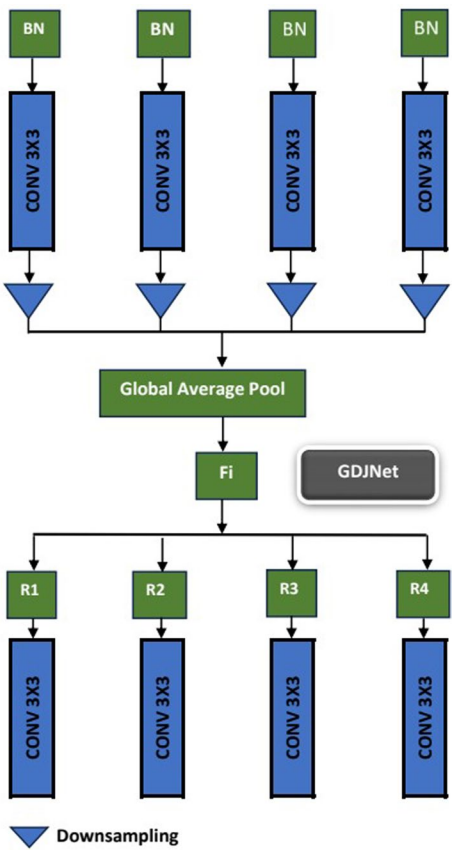

Fig. 2 Architecture of the proposed GDJNet

local features are extracted by applying the convolution operation. This helps the model to learn fine details and capture local patterns.

3.1.1.3 Feature fusion: integrating AAFAM and GDJNet To achieve effective feature representation, the feature maps extracted by AAFAM and GDJNet are fused. Channel-wise concatenation is employed to preserve fine-grained spatial details and global context. The final fused feature representation is given by:

$$FusedFeatures(F) = A_m(F1) \parallel G_m(F2). \tag{1}$$

The feature maps extracted by AAFAM and GDJNet is represented as $A_m(F1)$ and $G_m(F2)$, respectively. $\parallel$ indicates channel-wise concatenation.

3.1.1.4 Attention-guided fusion for anomaly enhancement Spatial attention improves anomaly localization by assigning higher weights to related regions. Channel attention makes a feature more discriminative by emphasizing those channels which are informative. These mechanisms are applied interchangeably to refine feature maps. In the same way, GDJNet estimates density distributions to model spatial relationships between objects. The element-wise multiplication is used for the merging of the feature maps:

$$Fused\ attention F_{attn} = D_{mp} \odot A_{attn}. \tag{2}$$

The density map generated by GDJNet is denoted as D_{mp}. A_{attn} is the attention map from AAFAM. $\odot$ represents element-wise multiplication. This fusion ensures a joint utilization of abnormality-aware and density-based cues. It reduces the effect of occlusions and background noise.

The resulting feature maps obtained after applying attention mechanisms undergo downsampling to reduce spatial dimensions. This preserves relevant motion patterns. The process here is to enhance the capability of anomaly localization by removing those parts of the background that are redundant. The proposed work's algorithm is given below:

```
# Input: Video frames (V)

Function Abnormality Aware Fused AttentionModel_GDJNet(V):

  # Step 1: Feature Extraction using Convolutional Layers
  For each frame f in V:
    - Apply Conv(3x3, 64) → (BN)
    - Apply Conv(1x1, 128) → Conv(3x3, 128) → BN
    - Apply Conv(1x1, 256) → Conv(3x3, 256) → BN
    - Apply Conv(1x1, 512) → Conv(3x3, 512) → BN

  # Step 2: Global Density Estimation via GDJNet
  - Density estimation using GDJNet
  - Apply GAP over entire feature map
  - Split feature map into four equal regions: (R1, R2, R3, R4)
  - Apply additional convolutional layers to extract fine-grained density features from each
    regions

  # Step 3: Attention Mechanisms for Feature Refinement
  - Apply SA to improve localization of abnormal regions
  - Apply CA to support feature discrimination

  # Step 4: Feature Fusion for Anomaly Enhancement
  - Concatenate AAFAM and GDJNet feature maps along the channel dimension:
    Fused Features (F) = A_m(F1) ‖ G_m(F2)                (1)
  - Compute fused attention representation:
    Fused Attention F_attn = D_mp ⊙ A_attn                (2)
  - Apply DS to reduce redundant spatial information
  - Use DC to refine feature maps for anomaly localization

  # Step 5: Final Classification
  - Apply Conv(3x3, 64) → Conv(1x1) layer
  - Output: Unusual Activity score

End Function
```

Algorithm for the proposed AAFAM–GDJNet.

3.2 Robustness under challenging conditions

The proposed framework shows a promising result in challenging conditions like occlusion, weak supervision, and low light. It applies the spatial and channel attention from AAFAM, as well as the local and global density modeling from GDJNet, so that the model can thoroughly detect the hidden anomalies. This allows the model to efficiently identify hidden anomalies. The attention-guided fusion mechanism highlights the most relevant regions where abnormalities are likely to occur. In low-light scenes, attention reduces noise and focuses on significant patterns. Together, these features enable the model to identify normal and unusual events.

4 Experimental results

4.1 Datasets

The different datasets such as the UMN dataset [38], the UCSD Ped2 dataset [39], the CUHK Avenue dataset, and the ShanghaiTech dataset have been used to analyze the efficiency of the proposed method. The UMN dataset is made up of 100 videos of the crowd on a college campus, in which anomalies like bicycles crossing pedestrian areas are present. The UCSD Ped2 dataset serves as a source for human trajectory prediction. Ped1 has 35 sequences with 15–20 people each, while Ped2 has 20 sequences with 50–100 people. The CUHK Avenue dataset consists of 16 training and 21 testing videos recorded from a fixed outdoor camera, this having anomalies such as running, loitering, and throwing objects. The ShanghaiTech dataset is a large-scale benchmark with 13 different scenes, offering varied abnormal events in complex environments.

We carried out an experiment with the division of data into training and testing sets in the ratio of 80:20, whereby 80% of the data was used for the training and 20% for the testing. There was no preprocessing done on the datasets. The performance metrics used were precision, recall, F1 score, accuracy, EER (equal error rates), and AUC (area under the curve).

4.2 Ablation study

The ablation study is conducted to measure the individual contributions of the AAFAM and the GDJNet. The results are presented in Table 1. The baseline model, without both AAFAM and GDJNet, achieves an F1-score of 88.0%. When only AAFAM is employed, the F1-score improves to 91.1%. This specifies its effectiveness in highlighting abnormal regions. Similarly, GDJNet alone results in an F1-score of 91.8%, showing its ability to capture both local and global spatial relationships. When combining both AAFAM and GDJNet, it achieves the highest F1-score of 93.2%. These results are consistent with recent advances in activity recognition, where approaches such as attention-based

Table 1 Evaluating the contribution of AAFAM and GDJNet

Model variant	Precision (%)	Recall (%)	F1 score (%)
Baseline model (without AAFAM and GDJNet)	89.2	86.8	88.0
AAFAM only (without GDJNet)	92.5	89.7	91.1
GDJNet only (without AAFAM)	93.1	90.5	91.8
Proposed model (AAFAM + GDJNet)	**95.2**	**91.2**	**93.2**

feature fusion [31], spatio-temporal modeling [32, 33], and generative prediction techniques [34] have shown a strong ability to capture complex motion patterns and contextual relationships.

Figure 3 displays the different areas and spatial distributions visually, which were determined as the most appropriate regions. The original frames used as input are shown in the first column. Attention maps that locate the hottest regions in each frame are shown in the second column. The third column represents density maps that are used for the spatial visualization of the detected objects.

4.3 Performance evaluation using K-fold cross-validation

Table 2 outlines the performance comparison of the proposed AAFAM–GDJNet. The evaluation is based on a fourfold cross-validation, which splits the datasets into four subsets for both training and testing. Precision goes up to 93.5%, and AUC is 95.5% at K = 3 on the UMN dataset, which is a clear indication of the model's capability in abnormal event detection. Consequently, the highest F1 score of 94.3% and AUC of 96.3% at K = 2 are recorded for the UCSD Ped2 dataset. Moreover, the model is very effective in the CUHK Avenue dataset as well, where the F1 score is above 92%, and the AUC is 95.2% at K = 4. Even if the scenes in the ShanghaiTech dataset are more complicated, the model is still able to attain an F1 score of 90.6% and an AUC of 93.0% at K = 4.

4.4 Equal error rate (EER) analysis

Table 3 presents the EER values obtained for the proposed method. The lower EER values indicate the model's effectiveness in distinguishing abnormal and normal activities. The UCSD Ped2 dataset exhibits the lowest EER of 11.4%, demonstrating the model's strong performance. Meanwhile, the ShanghaiTech dataset has a slightly higher EER of 18.3% because of its complex and crowded scenes. Overall, the results indicate that the proposed AAFAM–GDJNet effectively maintains strong anomaly detection performance.

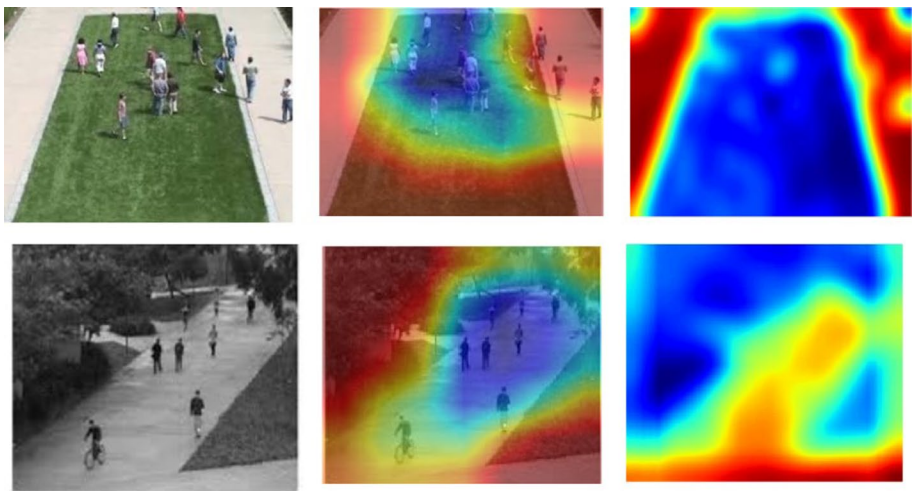

Fig. 3 First column represents the original frame, second column represents the attention map (highlights the most relevant regions), and the third column represents the density map (represents the spatial distribution of objects)

Table 2 Performance comparison of the proposed AAFAM–GDJNet model on different datasets

Dataset	K value	Precision (%)	Recall (%)	F1 score (%)	AUC (%)
UMN	1	92.3	90.5	91.4	95.2
	2	91.8	89.7	90.7	94.8
	3	93.5	91.2	92.3	95.5
	4	**92.7**	**90.9**	**91.8**	**95.0**
UCSD Ped2	1	94.8	92.6	93.7	96.1
	2	95.5	93.2	94.3	96.3
	3	94.3	92.9	93.6	96.0
	4	**94.7**	**93.3**	**94.0**	**96.2**
CUHK Avenue	1	93.7	91.4	92.5	94.6
	2	94.1	92.2	93.1	95.0
	3	93.9	91.9	92.9	94.8
	4	**94.3**	**92.5**	**93.4**	**95.2**
ShanghaiTech	1	90.9	88.7	89.8	92.3
	2	91.5	89.2	90.3	92.8
	3	91.2	88.9	90.0	92.5
	4	**91.8**	**89.5**	**90.6**	**93.0**

Table 3 Equal error rate (EER) analysis across datasets

Dataset	UMN	UCSD Ped2	CUHK Avenue	ShanghaiTech
EER (%)	12.7	11.4	14.1	18.3

Table 4 Comparative accuracy of proposed AAFAM–GDJNet with state-of-the-art techniques on benchmark datasets

	UMN	UCSD Ped2	CUHK Avenue	ShanghaiTech
Zhou et al. [40]	83.9	96.0	86.0	–
Tal et al. [41]		–	–	83.1
Hyun et al. [42]		97.2	86.8	74.0
Yang et al. [43]	–	97.9	84.2	83.8
Or Hirschorn et al. [44]	–	–	–	85.9
Micorek et al. [45]		–	94.03	85.9
Erkan et al. [46]	95.24	–	93.66	–
AAFAM–GDJNet	95.5	96.05	94.07	90.5

4.5 Accuracy comparison with state-of-the-art methods

Table 4 presents a comparative analysis of state-of-the-art techniques in terms of accuracy (%). The new model has achieved the highest accuracy compared to other methods on the UMN dataset (95.5%), and its results are better than those of Zhou et al. (83.9%) and Erkan et al. (95.24%). The AAFAM–GDJNet model with the UCSD Ped2 dataset recorded an accuracy of 96.05%, which is very close to the results of Yang et al. (97.9%) and Hyun et al. (97.2%). For the CUHK Avenue dataset, the AAFAM–GDJNet method, out of all the previous ones, with a 94.07% accuracy was the best. In the case of the ShanghaiTech dataset, the performance of AAFAM–GDJNet was 90.5%, hence having better results than Or Hirschorn et al. (85.9%) and Yang et al.

(83.8%). Generally, AAFAM–GDJNet most of the time holds very high accuracy on different datasets. AAFAM–GDN obtains the highest AUC score of 95.2%, as shown in Fig. 4, and it is superior to all other methods that were compared.

The computational efficiency of the proposed AAFAM–GDJNet model was assessed in terms of training time, inference time, and inference time per frame on two benchmark datasets: UMN and UCSD Ped2. On the UMN dataset, training took 120.5 s. The inference took 25.3 s and required an average inference time of 8.4 ms per frame. For the UCSD Ped2 dataset, training was completed in 95.8 s, and inference took 18.6 s. It requires 6.9 ms per frame. These results suggest that, despite a moderate training duration, the model delivers fast inference, making it well-suited for real-time applications.

5 Conclusion

The purpose of this research was to create an abnormal event detection system by coupling AAFAM with GDJNet. The new system was tested on four standard datasets: UMN and UCSD (Ped2), CUHK Avenue, and ShanghaiTech. The outcome showed that this fusion had a very significant impact on the detection of abnormal events. In AAFAM, attention mechanisms are used to direct the focus of the model to the abnormal regions of the feature maps. This allows the model to detect very weak anomalies and, at the same time, disregards the irrelevant information. By capturing both global and local spatial relationships, GDJNet can help AAFAM. The proposed work can understand the interactions between the objects at different scales. The abnormality-aware features needed to be combined with the global context to achieve better detection performance. The combination made it possible for the system to locate local anomalies and to understand the overall contextual patterns that resulted in higher accuracy and more robust abnormal event detection.

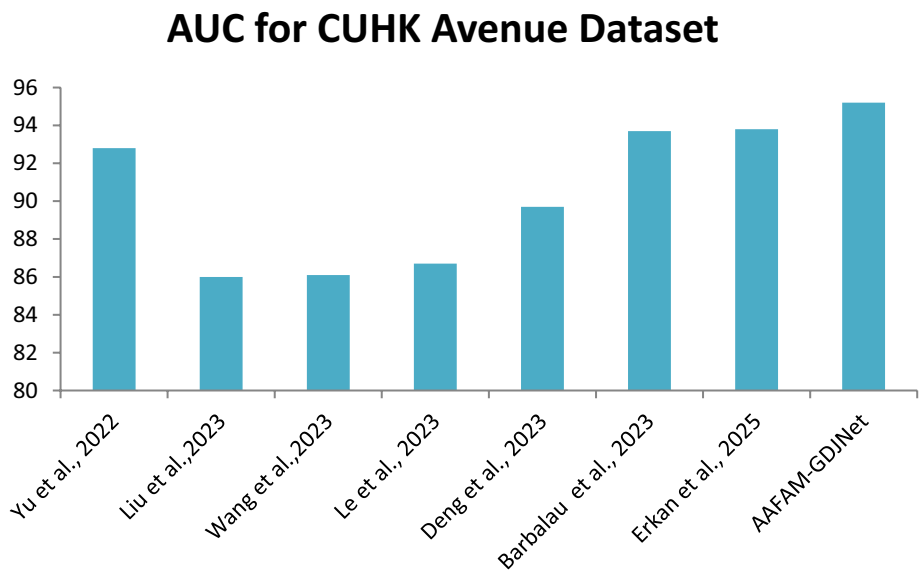

Fig. 4 Comparative AUC performance of different abnormal event detection methods

Author contributions
DSS: conceptualization, methodology design, supervision, and manuscript writing. RJV: implementation of the revised methodology, experimental analysis, and results validation. AAB: experimental analysis and critical review of the manuscript. SM: literature review, statistical analysis, and manuscript editing.

Funding
Not applicable.

Availability of data and materials
The datasets used in this research are publically available. UMN dataset—https://www.crcv.ucf.edu/projects/Abnormal_Crowd/, UCSD Ped2 dataset—http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Received: 22 August 2024 Accepted: 22 October 2025
Published online: 18 November 2025

References

1. M.A.Y. Peer Mohamed Appa, V. Vanitha, P. Rishi, S. Sagar, M.A. Paul, Key frame extraction based abnormal vehicle identification technique using statistical distribution analysis. Sci. Rep. **15**(1), 30957 (2025)
2. V. Saligrama, Z. Chen, Video anomaly detection based on local statistical aggregates, in *CVPR* (2012)
3. M.K. Asha Paul, J. Kavitha, P.A. Jansi Rani, Keyframe extraction techniques: a review. Recent Patents Comput. Sci. **11**(1), 3–16 (2018)
4. Y. Cong, J. Yuan, J. Liu, Sparse reconstruction cost for abnormal event detection, in *CVPR* (2011)
5. M. Asha Paul, P.A. Jansi Rani, L. Liba Manopriya, Gradient based aura feature extraction for coral reef classification. Wirel. Pers. Commun. **114**(1), 149–166 (2020)
6. M. Sabokrou, M. Fayyaz, M. Fathy, Z. Moayed, R. Klette, Deepanomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. Comput. Vis. Image Underst. **172**, 88–97 (2018)
7. X. Hu, Y. Huang, Q. Duan, W. Ci, J. Dai, H. Yang, Abnormal event detection in crowded scenes using histogram of oriented contextual gradient descriptor. EURASIP J. Adv. Signal Process. **2018**, 1–15 (2018)
8. E. Murali, A.S. Sheela, M.A. Paul, V. Muthu, A.Y. Felix. Learning normal patterns via conv-LSTM for video anomaly detection using likelihood statistical texture feature representation in surveillance videos. Int. J. Syst. Assur. Eng. Manag., pp.1–12 (2025)
9. C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **43**(4), 1445–1451 (2020)
10. J.E.D. Sheela, P.A. Jansi Rani, M.A. Paul, Super pixels transmission map-based object detection using deep neural network in UAV video. Imaging Sci. J. **71**(8), 767–775 (2023)
11. C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, X. Gao, Task-adaptive attention for image captioning. IEEE Trans. Circuits Syst. Video Technol. **32**(1), 43–51 (2021)
12. G. Chen, P. Liu, Z. Liu, H. Tang, L. Hong, J. Dong, J. Conradt, A. Knoll, Neuroaed: towards efficient abnormal event detection in visual surveillance with neuromorphic vision sensor. IEEE Trans. Inf. Forensics Secur. **16**, 923–936 (2020)
13. M.A. Paul, K.S. Kumar, S. Sagar, S. Sreeji, LWDS: lightweight deepseagrass technique for classifying seagrass from underwater images. Environ. Monit. Assess. **195**(5), 614 (2023)
14. A. Adam, E. Rivlin, I. Shimshoni, D. Reinitz, Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans. Pattern Anal. Mach. Intell. **30**(3), 555–560 (2008)
15. M. Javaid, A. Haleem, R.P. Singh, R. Suman, S. Rab, Significance of machine learning in healthcare: features, pillars and applications. Int. J. Intell. Netw. **3**, 58–73 (2022)
16. S.A. Jebur, K.A. Hussein, H.K. Hoomod, L. Alzubaidi, J. Santamaría, Review on deep learning approaches for anomaly event detection in video surveillance. Electronics **12**(1), 29 (2022)
17. J. Kavitha, P.A.J. Rani, P.M. Fathimal, A. Paul, An efficient shot boundary detection using data-cube searching technique. Recent Adv. Comput. Sci. Commun. **13**(4), 798–807 (2020)
18. M.A. Paul, P.A.J. Rani, Statistical modeling based directional pattern design (SMDPD) feature extraction for coral reef classification. Environ. Monit. Assess. **193**(9), 583 (2021)
19. Z. Zhang, L. Li, G. Cong, H. Yin, Y. Gao, C. Yan, A.V. Hengel, Y. Qi. From speaker to dubber: movie dubbing with prosody and duration consistency learning. ACM Multimedia (MM), (2024), pp.7523–7532

20. S. Bouindour, R. Hu, H. Snoussi. Enhanced convolutional neural network for abnormal event detection in video streams, in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, IEEE, (2019), pp. 172–178
21. V.K. Sharma, R.N. Mir, C. Singh, Scale-aware CNN for crowd density estimation and crowd behavior analysis. Comput. Electr. Eng. **106**, 108569 (2023). https://doi.org/10.1016/j.compeleceng.2022.108569
22. A. Patwal, M. Diwakar, V. Tripathi, P. Singh, An investigation of videos for abnormal behavior detection. Procedia Comput. Sci. **218**, 2264–2272 (2023)
23. M. Qasim, E. Verdu, Video anomaly detection system using deep convolutional and recurrent models. Results Eng. **18**, 101026 (2023)
24. G. Tang, H. Zhao, B. Yu, Low-cost and high-performance abnormal trajectory detection based on the GRU model with deep spatiotemporal sequence analysis in cloud computing. J. Cloud Comput. **13**(1), 53 (2024)
25. J. Amin, M.A. Anjum, K. Ibrar, M. Sharif, S. Kadry, R.G. Crespo, Detection of anomaly in surveillance videos using quantum convolutional neural networks. Image Vis. Comput. **135**, 104710 (2023). https://doi.org/10.1016/j.imavis.2023.104710
26. E.M. Onyema, S. Balasubaramanian, K. Suguna S, C. Iwendi, B.V.V.S. Prasad, C.D. Edeh, Remote monitoring system using slow-fast deep convolution neural network model for identifying antisocial activities in surveillance applications. Meas. Sensors **27**, 100718 (2023). https://doi.org/10.1016/j.measen.2023.100718
27. Y.S Chong, Y.H. Tay. Abnormal event detection in videos using spatiotemporal autoencoder, in *Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21–26, 2017, Proceedings, Part II 14*. Springer International Publishing, pp. 189–196
28. Ionescu RT, Khan FS, Georgescu MI, Shao L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), pp. 7842–7851
29. M. Asad, J. Yang, E. Tu, L. Chen, X. He, Anomaly3d: video anomaly detection based on 3d-normality clusters. J. Vis. Commun. Image Represent. **75**, 103047 (2021). https://doi.org/10.1016/j.jvcir.2021.103047
30. K.V. Thakare, D.P. Dogra, H. Choi, H. Kim, I.-J. Kim, Rareanom: a benchmark video dataset for rare type anomalies. Pattern Recognit. **140**, 109567 (2023). https://doi.org/10.1016/j.patcog.2023.109567
31. Z. Xiao, H. Tong, Federated contrastive learning with feature-based distillation for human activity recognition. IEEE Trans. Comput. Soc. Syst. (2025). https://doi.org/10.1109/TCSS.2024.3510428
32. N. Kumar, N. Sukavanam, An improved CNN framework for detecting and tracking human body in unconstraint environment. Knowl.-Based Syst. **193**, 105198 (2020)
33. N. Kumar, N. Sukavanam, Weakly supervised deep network for spatiotemporal localization and detection of human actions in wild conditions. Vis. Comput. **36**(9), 1809–1821 (2020)
34. H. Oğul, Language of actions: a generative model for activity recognition and next move prediction from motion sensors. Expert Syst. Appl. **264**, 125947 (2025)
35. H. Song, C. Sun, X. Wu, M. Chen, Y. Jia, Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos. IEEE Trans. Multimedia **22**(8), 2138–2148 (2019)
36. A. Vaswani, S. Noam, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf Process Syst. **30** (2017). https://doi.org/10.48550/arXiv.1706.03762
37. H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in *Proc. Int. Conf. Mach. Learn.*, (2019), pp. 7354–7363
38. University of Minnesota [Internet]. Detection of unusual crowd activity; [cited 2025 February 24], 2006. http://mha.cs.umn.edu/proj_events.shtml
39. V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos. Anomaly detection in crowded scenes, in *Proceedings of the Computer Vision and Pattern Recognition; 2010 Jun 13–18; San Francisco, CA, USA*, IEEE, New York, (2010)
40. J.T. Zhou, L. Zhang, Z. Fang, J. Du, X. Peng, Y. Xiao, Attention-driven loss for anomaly detection in video surveillance. IEEE Trans. Circuits Syst. Video Technol. **30**(12), 4639–4647 (2019)
41. Reiss T, Hoshen Y. Attribute-based representations for accurate and interpretable video anomaly detection. arXiv preprint arXiv:2212.00789, (2022)
42. W. Hyun, W.J. Nam, S.W. Lee, Dissimilate-and-assimilate strategy for video anomaly detection and localization. Neurocomputing **522**, 203–213 (2023)
43. Y. Yang, Z. Fu, S.M. Naqvi, Abnormal event detection for video surveillance using an enhanced two-stream fusion method. Neurocomputing **553**, 126561 (2023)
44. O. Hirschorn, S. Avidan. Normalizing flows for human pose anomaly detection, in *Proc. of the International Conference on Computer Vision (ICCV)*, (2023)
45. J. Micorek, H. Possegger, D. Narnhofer, H. Bischof, M. Kozinski. Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 18868–18877
46. B. Erkan, R. Samet, Q. Al-Haija, A. Alqahtani, R. A. Alsemmeari, B. Alghamdi, B. Alturki, A. Alsulami, Abnormal event detection in surveillance videos through LSTM auto-encoding and local minima assistance. Discov Internet of Things **5**(1), 32 (2025)
47. Q. Li, R. Yang, F. Xiao, B. Bhanu, F. Zhang, Attention-based anomaly detection in multi-view surveillance videos. Knowl. Based Syst. **252**, 109348 (2022)
48. G. Yu, S. Wang, Z. Cai, X. Liu, C. Xu, C. Wu. Deep anomaly discovery from unlabeled videos via normality advantage and self-paced refinement, in *Proceedings of the 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*; IEEE; (2022). p. 13967–78.
49. Y. Liu, Z. Guo, J. Liu, C. Li, L. Song, Osin: object-centric scene inference network for unsupervised video anomaly detection. IEEE Signal Process. Lett. **30**, 359–363 (2023)
50. Y. Wang, T. Liu, J. Zhou, J. Guan, Video anomaly detection based on spatio-temporal relationships among objects. Neurocomputing **532**, 141–151 (2023)

51. V.-T. Le, Y.-G. Kim, Attention-based residual autoencoder for video anomaly detection. Appl. Intell. **53**, 3240–3254 (2023)
52. H. Deng, Z. Zhang, S. Zou, X. Li. Bi-directional frame interpolation for unsupervised video anomaly detection, in *Proceedings of the 2023 IEEE/CVF winter conference on applications of computer vision (WACV)*; IEEE, (2023), p. 2633–42
53. A. Barbalau, R.T. Ionescu, M.-I. Georgescu, J. Dueholm, B. Ramachandra, K. Nasrollahi, F.S. Khan, T.B. Moeslund, M. Shah, SSMTL++: revisiting self-supervised multi-task learning for video anomaly detection. Comput. Vis. Image Underst. **229**, 103656 (2023)

Publisher's Note