IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 5, October, 2025, pp. 2609-2618
Journal homepage: http://iieta.org/journals/ts

Automatic Early Warning and Visual Analysis Framework for Sudden Health Incidents in Public Spaces Based on Multi-Source Video Streams and Behavior Recognition Algorithms

Sizhe Fan[1], Zihan Mu[2*]

[1] Transportation Institute of Inner Mongolia University, Hohhot City 010030, China
[2] Shandong University-Australian National University Joint College of Science, Shandong University (Weihai), Weihai 264209, China

Corresponding Author Email: 202300700044@mail.sdu.edu.cn

https://doi.org/10.18280/ts.420514

Received: 1 March 2025
Revised: 7 August 2025
Accepted: 22 August 2025
Available online: 31 October 2025

Keywords:
public spaces, sudden health incidents, multi-source video streams, behavior recognition, automatic early warning, visual analysis, object tracking, cross-frame attention module

ABSTRACT

With the acceleration of urbanization, the density of people in public spaces such as airports and shopping malls continues to rise. Sudden health incidents, such as fainting and acute cardiovascular events, pose a serious threat to public safety due to their sudden nature and short rescue window. Traditional manual monitoring is limited by labor costs and attention decay, making real-time and accurate early warning of such incidents difficult. The growing demand for intelligent public safety governance, coupled with advancements in artificial intelligence technologies, provides an opportunity for technological innovation in this field. Although existing video surveillance and behavior recognition technologies have been applied in public safety, they still face significant challenges in complex scenarios: crowd density and occlusion reduce the robustness of object detection and tracking, there is a lack of synchronization and fusion capabilities for multi-source video streams, the early warning system and visual analysis are disconnected, and the quality of specialized datasets limits algorithm optimization. To address these issues, this paper proposes an automatic early warning and visual analysis framework for sudden health incidents in public spaces, based on multi-source video streams and behavior recognition algorithms, and builds an end-to-end intelligent analysis pipeline. The framework achieves multi-source video stream synchronization through a distributed access mechanism, and extracts multi-scale features using the DarkNet53 backbone network. An innovative cross-frame attention module is introduced to strengthen the target area expression by associating temporal features, improving detection stability in occlusion and deformation scenarios. A twin network is used to achieve precise target displacement prediction, and a data association strategy based on appearance similarity and motion consistency is employed to ensure the continuity and identity consistency of individual spatiotemporal trajectories. Finally, relying on multi-source information fusion, the framework triggers multi-level automatic early warnings, and constructs a visual analysis platform with a Web GIS map, data dashboard, and interpretable video summaries to support emergency decision-making. This research provides a technical paradigm and practical solution for the intelligent management of sudden health incidents in public spaces, while enriching the theoretical and methodological applications of computer vision in the field of public safety.

1. INTRODUCTION

With the acceleration of urbanization [1, 2], the flow of people in public spaces such as airports, shopping malls, and stations continues to rise, and spatial scenes are becoming increasingly complex, significantly increasing the risk of sudden health incidents [3]. These incidents are characterized by their strong suddenness, rapid progression, and short rescue window. If they are not detected in time and emergency responses are not initiated promptly, they can easily lead to serious life safety accidents, which in turn affect the stability of social order [4, 5]. The traditional management model relying on manual monitoring is limited by high labor costs, attention decay, and low efficiency in multi-region collaboration, making it difficult to achieve real-time awareness and accurate early warning of sudden health incidents in large-scale areas [6, 7]. Meanwhile, the demand for intelligent public safety governance [8] is becoming increasingly urgent, and the rapid development of artificial intelligence and computer vision technologies [9, 10] provides a technical possibility for solving this problem. Constructing an efficient intelligent early warning system has become a core requirement in the field of public space safety management.

Research on automatic early warning and analysis of sudden health incidents in public spaces has important theoretical and practical significance. In practice, this research can break through the limitations of traditional manual monitoring and realize the early identification, rapid early warning, and visual

traceability of sudden health incidents, providing critical time for emergency rescue, effectively reducing casualties and property losses, and helping improve the fine and intelligent level of public space safety governance. It aligns with the construction needs of intelligent multi-point triggering monitoring and early warning systems. In theory, this research focuses on the challenges of target perception and behavior understanding in complex scenarios such as dense crowds and frequent occlusions, exploring synchronization and fusion of multi-source video streams and temporal feature extraction technologies. It can enrich the application results of computer vision in the field of public safety and provide technical references and methodological support for anomaly event recognition research in similar complex scenarios.

Although existing related research makes certain progress, there are still many technical deficiencies and application bottlenecks. First, the robustness of object detection and tracking is insufficient. Pervaiz et al. [11] pointed out in a review of abnormal behavior recognition of pedestrians in public places that existing methods perform poorly in scenes with dense crowds and frequent occlusions. Traditional methods based on background subtraction [12] are sensitive to scene complexity and lighting changes, while appearance-based methods face challenges in feature extraction. At the same time, related studies show that most trackers use short-term memory mechanisms and tend to lose targets in long-term occlusion scenarios, resulting in tracking interruptions [13-15]. Second, there is a lack of multi-source information fusion capabilities. Existing solutions often fail to effectively address the synchronization issue of multi-perspective video streams, leading to large matching errors across views and difficulty in forming globally consistent trajectory information, thus affecting the accuracy of event judgment [16, 17]. Third, the early warning and analysis stages are disconnected. Traditional early warning systems not only have a single data source and outdated technologies but also lack intelligent learning capabilities. Moreover, most behavior recognition algorithms do not have well-developed visual analysis modules, which are unable to provide intuitive support for emergency decision-making, resulting in low response efficiency [18, 19]. Fourth, the quality of specialized datasets is insufficient. The anomaly behavior recognition field lacks high-quality datasets with a rich variety of behaviors and clear definitions for crowd abnormal behavior, which restricts the optimization and verification of algorithm performance [4, 20].

To address the aforementioned research shortcomings, this paper proposes an automatic early warning and visual analysis framework for sudden health incidents in public spaces based on multi-source video streams and behavior recognition algorithms, constructing an end-to-end intelligent analysis pipeline from multi-perspective perception to behavior understanding. The core components of the framework include: using a distributed access mechanism to achieve synchronized collection of multi-source video streams; designing an end-to-end multi-object tracking network and innovatively introducing a cross-frame attention module to enhance temporal feature expression, improving detection stability in occlusion and deformation scenarios; combining twin networks for target displacement prediction and appearance-motion multi-dimensional data association strategies to achieve stable tracking of individual spatiotemporal trajectories and identity consistency; building a multi-source information fusion center and a visual analysis platform to trigger multi-level automatic early warnings and generate interpretable video summaries. The core value of this research lies in: theoretically breaking through the performance bottlenecks of traditional algorithms in complex scenarios, improving the technical path for temporal feature utilization and multi-source data fusion; and practically achieving a seamless connection from precise identification of sudden health incidents to efficient emergency response, providing a practical and operable intelligent solution for public space safety management.

2. METHODS

2.1 Algorithm working principle and network model design

In order to build an end-to-end intelligent analysis pipeline from multi-perspective perception to behavior understanding, the multi-source video stream and behavior recognition algorithm for automatic early warning of sudden health incidents in public spaces is proposed in this paper. The framework first synchronously acquires monitoring views from different spatial locations through a distributed multi-source video stream access layer. Each video stream independently enters a carefully designed end-to-end multi-object tracking network, which uses the efficient DarkNet53 as the backbone network to extract rich multi-scale features. Given the characteristics of crowded spaces and frequent occlusions in public places, the algorithm innovatively introduces a cross-frame attention module. This module calculates the spatial correlation between the current frame and historical frame feature maps to generate an attention weight map, thereby enhancing the potential target area features that persist over time in the current frame. This temporal enhancement mechanism significantly improves detection robustness in cases of brief occlusion or deformation. Subsequently, the target displacement prediction module adopts a twin network structure, and through deep cross-correlation operations within the search area, it accurately regresses the target's position shift between consecutive frames. Furthermore, by integrating appearance similarity and motion consistency through a data association strategy, the system outputs stable, continuous, and identity-consistent spatiotemporal trajectories for each individual. The specific framework structure is shown in Figure 1.

The closed-loop is completed through the integrated automatic early warning and visual analysis module. Once the multi-source information fusion center confirms the event, the system instantly triggers multi-level early warnings via an efficient message queue, pushing structured alarm information that includes event type, precise GIS location, timestamp, and associated video links to the command center and mobile terminals. Meanwhile, the visual platform is launched synchronously, highlighting event points on the Web GIS map in real-time, dynamically updating the global situation on the data dashboard, and automatically generating interpretable video summaries overlaid with skeletal key point trajectories and behavior labels. This provides intuitive decision support for security personnel, thus achieving a seamless connection from accurate perception to efficient emergency response.

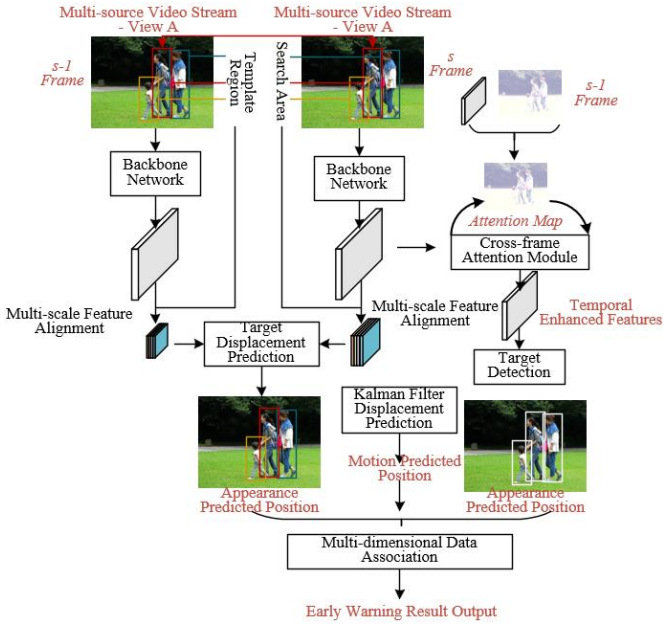

Figure 1. Public space sudden health incident early warning framework based on multi-source video streams and behavior recognition

2.2 Object detection

To address the challenges of varying target scales, dense distribution, and high real-time detection requirements in public space surveillance scenarios, this paper designs an anchor-free object detection module. The traditional anchor-based mechanism, due to its predefined shapes and numbers, lacks generalization capability when dealing with individuals with varying postures in sudden health incidents, and the hyperparameter tuning is complex. Therefore, we adopt and deepen the anchor-free paradigm of YOLOX, using it as the basis for target perception in our multi-source video stream analysis framework. This module maps the input image directly to dense anchors and lets each anchor directly regress the center offset and width-height of the target box, greatly simplifying the detection process and significantly improving computational efficiency. In particular, from a macro perspective of video sequence analysis, we introduce a cross-frame attention enhancement mechanism for the regression branch feature map. This mechanism dynamically strengthens the feature responses that persist over time and potentially represent targets at risk, by calculating the spatial cross-correlation between the current frame and historical frame feature maps. This allows the detector to not only rely on the appearance information of a single frame but also utilize temporal context, effectively suppressing missed detections caused by brief occlusions or motion blur, ensuring continuous and stable localization for each individual who may experience a health incident, and providing reliable input for subsequent tracking and behavior recognition. Figure 2 shows the network architecture of the object detection module.

To detect targets from small objects in the distance to large objects in close-up, and precisely capture individuals in public spaces who are either standing or lying down, this paper constructs an efficient multi-scale feature pyramid architecture. The backbone network uses DarkNet-53, which strikes an excellent balance between speed and accuracy, and cascades the Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) to build a multi-scale feature pyramid rich in semantic information and high-resolution details. The detection head performs parallel predictions on feature maps at three different scales, with each scale prediction using a decoupled design, where the classification task and regression task are separated and handled by independent convolutional branches. This decoupled structure avoids the interference of classification and regression tasks at the feature level, allowing the network to focus more on each task and significantly improving the box localization accuracy. Finally, the module outputs the target boxes after Non-Maximum Suppression (NMS), which have high confidence and are seamlessly passed to the subsequent cross-frame attention twin tracking module, forming a complete end-to-end link from accurate perception to stable tracking. Figure 2 provides the network architecture of the object detection module.

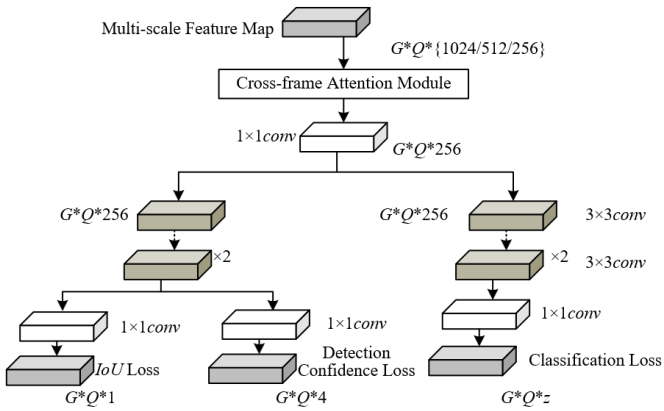

Figure 2. Network architecture of the object detection module

2.3 Target displacement prediction module

To ensure stable and continuous tracking of potential patient targets in complex public spaces, this paper designs a target displacement prediction module based on a twin network to solve the association drift problem that arises in scenarios with dense crowds, frequent occlusions, and visually similar targets. The core of this module lies in the use of a spatiotemporal matching mechanism based on appearance similarity: the target image area confirmed in the previous frame is used as a template, and in the current frame, a larger search area is constructed around its historical position. Through a twin network with shared parameters, deep features of both the template and the search area are extracted. Then, by performing channel-wise correlation operations, a response feature map is generated, where each peak in the spatial domain represents the candidate position in the search area most similar in appearance to the template target. Using the aforementioned spatiotemporal matching mechanism, the module allows the system to perform "recognition" across consecutive frames based on the target's own visual features, without relying on drastic motion assumptions. This approach can effectively handle non-rigid deformations and brief full occlusions that may occur in individuals during sudden health events, ensuring that the trajectory does not break due to sudden failure of the motion model. Figure 3 illustrates the network architecture of the target displacement prediction module.

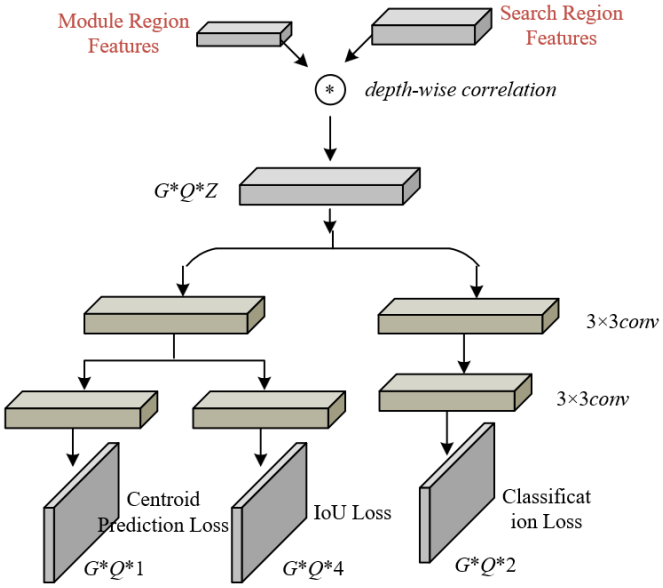

Figure 3. Network architecture of the target displacement prediction module

In the specific design of the network model, we adopt an anchor-free prediction mechanism aimed at precise regression, to accommodate the dramatic changes in target scale and posture in public spaces. Starting from the response feature map obtained from the correlation operations, we construct three parallel prediction branches: the foreground-background classification branch, the bounding box regression branch, and the center-ness prediction branch. The foreground-background classification branch is responsible for determining whether each point in the response feature map belongs to the real target foreground or background interference, with its ground truth set within an area half the width and height of the target's true bounding box, thus defining a compact positive sample area. The bounding box regression branch directly predicts the distances of the target box's four boundaries (top, bottom, left, right) from the current anchor point. This anchor-free parameterization avoids the issue of mismatch between predefined anchor boxes and the true shape of targets, which is particularly suitable for accurately locating individuals with irregular postures and abnormal aspect ratios, such as those who faint or fall. To further improve the localization quality, the center-ness prediction branch is introduced to assess the reliability of each foreground point's prediction result—the closer the point is to the target's geometric center, the more accurate the predicted bounding box is. Assuming the distance from the point to the four boundaries of the true bounding box is represented by s, y, m, and e, the ground truth of center-ness can be computed based on the point's position within the box:

$$CE = \sqrt{\frac{MIN(m,e)}{MAX(m,e)} * \frac{MIN(s,y)}{MAX(s,y)}} \tag{1}$$

The output of this branch serves as a soft weight, which is multiplied by the classification confidence during inference, thus suppressing unstable predictions that are classified as foreground but located at the edges of the target. This results in more accurate and stable displacement prediction boxes from the system.

Finally, the displacement prediction module outputs the unique and optimal target position estimate through a carefully designed decision fusion and penalty mechanism, providing high-fidelity trajectory data for the entire early warning system. During inference, the network integrates foreground-background classification prediction, center-ness prediction, and introduces cosine window penalties for temporal consistency and scale variation penalties for morphological stability, collectively generating the final confidence feature map. Assuming a point on the feature map is represented by a, the center of the feature map is represented by a_z, and the size of the feature map is represented by V, the cosine window penalty for the target's position change is calculated as:

$$o_z = 0.5 - 0.5 * COS\left(\frac{2\tau * DIS(a, a_z)}{(V-1)/2 - 1}\right) \tag{2}$$

Assuming the aspect ratio of the previous and current time is represented by e and e', and the scale at the previous and current time by t and t', the element-wise multiplication for the same location is represented by $\otimes$, and the hyperparameter is represented by j, the scale variation penalty mechanism is calculated as:

$$o_t = r^{j*MAX\left(\frac{e}{e'}, \frac{e'}{e}\right) \otimes MAX\left(\frac{t}{t'}, \frac{t'}{t}\right)} \tag{3}$$

By selecting the point with the highest confidence in the feature map, the corresponding bounding box regression output is the target displacement prediction result for the current frame. In summary, based on the foreground-background classification prediction feature map M_{CLS}, center-ness prediction feature map M_{CTN}, and the penalties for position and scale variation o_z, o_t, with the hyperparameter λ controlling the strength of the cosine window penalty, the final confidence feature map t can be obtained:

$$t = M_{CLS} \otimes M_{CTN} \otimes o_t * (1-\lambda) + o_z * \lambda \tag{4}$$

Through the multi-factor decision process, this module greatly enhances the system's robustness in tracking specific individuals in a mixed crowd environment, effectively resisting interference from similar-looking individuals and its own appearance mutations. The continuous and accurate trajectory generated from this process is the fundamental prerequisite for the subsequent behavior recognition module to reliably detect key patterns of sudden health events, such as "sudden collapse" and "prolonged stillness."

2.4 Attention mechanism

In scenarios with high crowd density and complex flow dynamics, such as subway stations and plazas, individuals who suddenly faint or fall are easily completely obscured by other pedestrians in a short period of time. Traditional detectors, in this case, will miss detection due to the temporary loss of visual information about the target, leading to interrupted tracking and failure of the early warning system. To improve the robustness of multi-target tracking in the public health emergency early warning system, this paper designs a cross-frame attention module based on the temporal continuity of video data to solve the tracking interruption problem caused by occlusion, rapid motion, and appearance changes in complex public scenarios.

The core working principle of this module is to construct a spatiotemporal attention mechanism based on historical tracking results. It encodes the spatial prior information of the confirmed target in the previous frame as attention weights in the form of a Gaussian heatmap. The Gaussian heatmap is generated by radiating from the historical target position, clearly marking the probability distribution of the target's appearance in the previous frame. Subsequently, by calculating the spatial association matrix between the current frame and the historical frame feature maps, and using this matrix to transfer the attention weights of the historical frame to the current frame, an attention weight map predicting the possible region where the target might appear in the current frame is generated. The essence of the module is to model the spatiotemporal existence of the target as a conditional probability problem, enabling the network to actively and purposefully enhance the feature response of potential target areas in the current frame, significantly reducing the risk of missed detection due to brief full occlusion or target deformation. Figure 4 shows the cross-frame attention module network architecture.

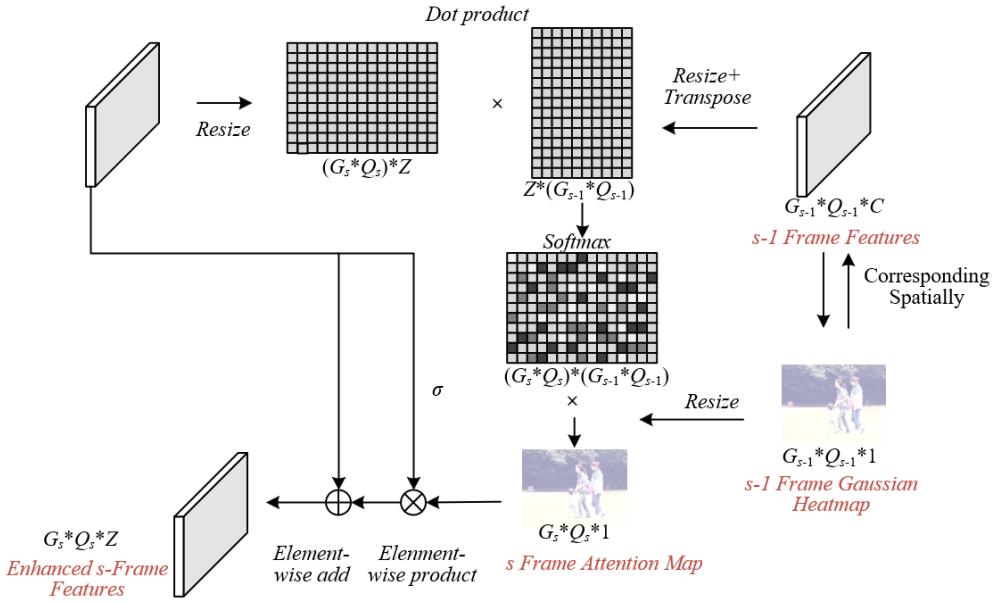

Figure 4. Cross-frame attention module network architecture

In the specific implementation of the network model, the module constructs a refined, differentiable information transmission pathway. First, the system uses the tracking result of the s-1 frame to generate a Gaussian heatmap HG_{s-1} with the same resolution as the feature map, where the value of each pixel represents the probability of that position being the foreground target. Specifically, suppose the set of center points of the target bounding boxes in the s-1 frame tracking result is represented by $\{o_0, o_1,\ldots\}$, and the radiation range coefficient for each target point is represented by δ_u. The value of a point w on the map is calculated as follows:

$$HG_{s-1}\left(w,\{o_0,o_1...\}\right)=\underset{u}{MAX}\left[\exp\left(-\frac{(o_u-w)^2}{2\delta_u^2}\right)\right] \tag{5}$$

Next, by calculating the dot product between the s frame feature map D_s and the s-1 frame feature map D_{s-1}, a spatial association matrix is obtained. This matrix is normalized by Softmax and represents the similarity of all spatial location pairs between the two frames. The association matrix and the Gaussian heatmap HG_{s-1} are then used for a dot product operation, which transfers the attention weights across frames, outputting the attention weight map XL_s for the current frame. This weight map accurately reflects the likelihood of the target appearing at various positions in the current frame based on historical information. Suppose the *resize* operation that merges the width and height dimensions of the feature map is represented by $E(\cdot)$. The above calculation process is expressed as:

$$XL_s=\text{softmax}\left(E(D_s)*E(D_{s-1})^T\right)*E(HG_{s-1}) \tag{6}$$

Finally, through a learnable scaling parameter σ, the attention weight map is fused with the original feature map in a feature enhancement manner, resulting in an enhanced feature map D_s'. Suppose the addition of corresponding positions in the feature maps is represented by $\oplus$, and XL_s is expanded to the same dimensions as D_s, represented by $\otimes$. The calculation formula is as follows:

$$D_s^{'}=D_s\oplus(\sigma*XL_s\otimes D_s) \tag{7}$$

The cross-frame attention module enables the network to adaptively adjust its dependence on historical information during training, neither blindly trusting nor easily discarding it, thereby achieving the best balance between tracking stability and adaptation to changes in target appearance.

2.5 Automatic early warning and visual analysis

To achieve a seamless closed-loop from algorithm recognition to on-site response, this paper constructs an automatic early warning and visual analysis module, which serves as the critical output layer of the entire system from perception computation to decision support. The module is designed based on a collaborative decision-making event trigger that fuses multi-source information. When the system determines that an individual's behavior sequence, through the front-end multi-target tracking and behavior recognition

network, matches a pre-set sudden health event model and the confidence exceeds the threshold after fusing evidence from multiple camera views, the trigger is activated. The system then asynchronously pushes structured early warning messages to multiple nodes such as security centers and mobile inspection terminals through a high-concurrency message queue. These messages are not just simple alarm signals but rather data packets that integrate event types, precise geographic locations, accurate timestamps, as well as associated target IDs and video clip indexes. Through a low-latency, structured communication mechanism, the module ensures that early warning information can be accurately and quickly distributed to the relevant personnel, gaining precious time for initiating emergency rescue.

In terms of network model and system design, this module builds a visual analysis platform that integrates WebGIS technology, real-time data stream processing, and video semantic enhancement. The platform first uses the WebGIS engine to register the dispersed cameras as perception nodes on a digital map. When the early warning is triggered, the corresponding location will immediately flash on the global situation map, achieving accurate geographic positioning of the event and global situational awareness. Simultaneously, a real-time data screen built based on libraries such as ECharts dynamically aggregates macro information for the entire location, such as real-time crowd density in each area, statistical distribution of different event types, and historical trends, thereby identifying risk hotspots and assisting in proactive scheduling of security resources. Crucially, the system integrates video summarization and explainable backtracking functionality. It automatically extracts key periods before and after the event from multi-angle video streams and generates a condensed summary. During playback, the system overlays the motion trajectory and behavior recognition labels generated by the front-end algorithm onto the original video. Through deep visual enhancement, the module transforms the algorithm's "black-box" judgments into an "evidence chain" that security personnel can intuitively understand, greatly enhancing the transparency and credibility of the system's decisions, allowing security personnel to quickly verify the situation and take the most appropriate action.

2.6 Loss function

The loss function designed in this paper aims to guide the entire model to precisely adapt to the stringent requirements of sudden health event early warning in public spaces through a multi-task collaborative optimization strategy. This loss function is constructed as the weighted sum of the target detection loss $LOSS_{JC}$ in the tracking algorithm and the target displacement prediction loss $LOSS_{YC}$. Among them, the target detection loss $LOSS_{JC}$ focuses on achieving high recall and precise localization of all pedestrians in a single-frame image. The classification cross-entropy loss $LOSS_{FL}$ included in this ensures that the model can effectively distinguish foreground targets from complex backgrounds, while the *IoU* loss $LOSS_{IoU}$ drives the predicted box to align closely with the ground truth box at the boundaries, ensuring the accurate analysis of the spatial relationship between the human body and the ground. The detection confidence loss $LOSS_{ZXD}$ further optimizes the model's ability to identify valid targets in dense crowds. The calculation formula for $LOSS_{JC}$ is:

$$LOSS_{JC} = LOSS_{FL} + LOSS_{IoU} + LOSS_{ZXD} \tag{8}$$

Assuming that the width and height of the feature map are represented by G and Q, the total number of detection categories is J, the ground truth category probability is represented by b_k, and the predicted category probability is represented by o', the formulas for $LOSS_{FL}$ and $LOSS_{ZXD}$ are:

$$LOSS_{FL} = -\frac{1}{G*Q}\sum_{k=1}^{J}\sum_{u=1}^{G*Q} b_k \log\left(o_u^{'}\right) \tag{9}$$

$$LOSS_{ZXD} = -\frac{1}{G*Q}\sum_{u=1}^{G*Q}\left[o_u \log\left(o_u^{'}\right) + \left(1-o_u\right)\log\left(1-o_u^{'}\right)\right] \tag{10}$$

Assuming that the ground truth bounding box and the predicted bounding box are represented by BOX_{GT} and BOX_{PR}, the detailed definition of $LOSS_{IoU}$ is:

$$LOSS_{IoU} = 1 - IoU\left(BOX_{GT}, BOX_{PR}\right) \tag{11}$$

The design of the target displacement prediction loss $LOSS_{YC}$ directly serves the key indicator of trajectory continuity and stability. Its foreground-background classification loss $LOSS_{FL1}$ drives the twin network to clearly distinguish between the target and visually similar interference objects within the search area, which is crucial for preventing tracking identity shifts in crowded scenarios. Similarly, the box regression loss $LOSS_{IoU1}$, in *IoU* form, ensures the geometric accuracy of the displacement prediction box. Especially important is the introduction of the center-ness prediction loss $LOSS_{ZXD1}$, which supervises the model through cross-entropy loss, encouraging it to trust prediction points closer to the target's geometric center, thereby outputting more stable and reliable bounding boxes. The loss function for the $LOSS_{YC}$ part is defined as follows:

$$LOSS_{YC} = LOSS_{FL1} + LOSS_{IoU1} + LOSS_{ZXD1} \tag{12}$$

The total loss function of the algorithm is:

$$LOSS = \lambda_1 * LOSS_{JC} + \lambda_2 * LOSS_{YC} \tag{13}$$

The use of the above loss function effectively suppresses frame jitter caused by target deformation or partial occlusion, thus providing smooth and continuous spatiotemporal trajectory data for the behavior recognition module, laying the foundation for the high reliability of the entire early warning system from the optimization objective layer.

3. EXPERIMENTS

To verify the enhancement of the cross-frame attention module on the robustness of target perception in crowded occlusion scenarios in public spaces, the ablation experiment was conducted. From the data in Table 1, it can be seen that without the cross-frame attention module, the performance in terms of Multi-Object Tracking Accuracy (MOTA) and target detection recall was the worst, with higher ID switch rates and false positive rates. This indicates that the lack of temporal feature correlation leads to insufficient stability in target tracking and detection in dense crowds and frequently occluded public spaces. As the history frame length of the cross-frame attention module increased, MOTA improved from 35.2 to 38.7, the ID switches/frame decreased from 0.85

to 0.35, target detection recall increased from 0.78 to 0.93, and the false positive rate decreased from 0.12 to 0.05. This demonstrates that the cross-frame attention module improves detection and tracking robustness during temporary occlusions or deformations by enhancing the feature correlation between the current frame and historical frames. Meanwhile, the processing frame rate slightly decreased as the history frame length increased but remained above 35 Hz, meeting the real-time early warning performance requirements in public spaces. In conclusion, the introduction of the cross-frame attention module significantly enhanced the perception capability of the multi-target tracking network in complex scenarios, providing a reliable target feature foundation for subsequent behavior recognition and event early warning.

Table 1. Ablation experiment results for cross-frame attention module

Experimental Setup	MOTA	ID Switches/Frame	Target Detection Recall	False Positive Rate	Processing Frame Rate (Hz)
No Cross-frame Attention	35.2	0.85	0.78	0.12	45
Cross-frame Attention (History Frame Length = 3)	37.8	0.52	0.89	0.08	42
Cross-frame Attention (History Frame Length = 5)	38.5	0.38	0.92	0.06	38
Cross-frame Attention (History Frame Length = 7)	38.7	0.35	0.93	0.05	35

Table 2. Performance experiment results of behavior recognition algorithm

Algorithm	Proposed Algorithm	*ST-GCN*	*I3D*	*Two-StreamCNN*
Fall Recognition Accuracy	0.92	0.85	0.88	0.86
Syncope Recognition Accuracy	0.90	0.82	0.85	0.83
Acute Cardiovascular Event Recognition Accuracy	0.88	0.78	0.82	0.80
F1 Score	0.90	0.81	0.85	0.83
Missed Detection Rate	0.05	0.12	0.09	0.10
False Positive Rate	0.06	0.15	0.10	0.12
Average Accuracy in Dense Crowds	0.89	0.80	0.84	0.82
Average Accuracy in Sparse Crowds	0.93	0.85	0.88	0.86
Average Accuracy in Strong Light	0.91	0.83	0.86	0.84
Average Accuracy in Weak Light	0.88	0.75	0.80	0.78

To validate the behavior recognition module's ability to accurately identify typical sudden health events in public spaces, the above performance comparison experiments were conducted. From the data in Table 2, it can be observed that the proposed algorithm significantly outperforms mainstream behavior recognition algorithms such as ST-GCN, I3D, and Two-StreamCNN in terms of recognition accuracy for various sudden health events and F1 score. Specifically, the fall recognition accuracy reaches 0.92, syncope recognition is at 0.90, and acute cardiovascular event recognition is at 0.88, with an F1 score of 0.90. The missed detection rate and false positive rate are as low as 0.05 and 0.06, respectively. In terms of scene robustness, the proposed algorithm maintains an average accuracy above 0.88 in dense crowds, sparse crowds, strong light, and weak light conditions, while the performance of the comparison algorithms clearly deteriorates in these scenarios. This indicates that the behavior recognition module proposed in this paper, by integrating temporal and spatial features from multi-source video streams, not only achieves high-precision recognition of various sudden health events but also exhibits strong scene adaptability, capable of handling complex situations such as fluctuations in crowd density and lighting changes in public spaces. This performance advantage ensures the early and accurate recognition of sudden health events, providing crucial decision support for subsequent automatic early warning and emergency response.

To verify the enhancement effect of the cross-frame attention module and multimodal feature fusion on backbone network performance, experiments were conducted to examine the relationship between multi-object tracking accuracy and processing frame rate under different backbone network configurations. From the data in Figure 5, it can be seen that without the cross-frame attention module, the MOTA remained between 35 and 38.5 when the frame rate was between 10 Hz and 50 Hz, showing some fluctuations but generally stable performance.

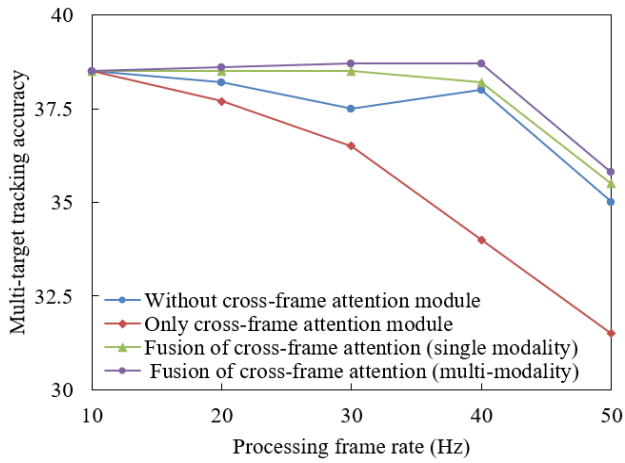

Figure 5. Relationship between multi-object tracking accuracy and processing frame rate under different backbone network configurations

With only the cross-frame attention module, performance decayed significantly as the frame rate increased, with MOTA reaching only 31.5 at 50 Hz, indicating that the introduction of cross-frame attention alone, without multi-scale feature support, struggles with the computational pressure at high frame rates. When cross-frame attention was combined with multimodal feature fusion, MOTA stabilized between 38.5

and 38.2 in the 30-40 Hz range, reaching a peak of 38.7 at 30-40 Hz, and maintaining a high level of 35.8 even at 50 Hz. This shows that the combination of the cross-frame attention module and multimodal feature fusion effectively strengthens temporal dimension target feature correlation and significantly improves tracking robustness in crowded, frequently occluded public space scenarios. This performance advantage provides a stable target localization and feature representation foundation for the subsequent behavior recognition module, ensuring continuous perception of individual behavior during sudden health events.

To verify the advantages of the proposed data association strategy in trajectory continuity and identity consistency in multi-source video stream scenarios, experiments were conducted to examine the relationship between multi-object tracking accuracy and processing frame rate under different data association strategies. Figure 6 compares the current mainstream tracking algorithms *ByteTrack*, *BoT-SORT*, *StrongSORT*, and the proposed fusion strategy, which shows significant performance advantages across the full frame rate range. At 10Hz, MOTA is 38, and at 50 Hz, it remains at 34.5. Although the performance at 60Hz has not fully shown its potential, the trend remains competitive. In contrast, *ByteTrack's* MOTA drops to 34 at 50 Hz, and *BoT-SORT* and *StrongSORT* degrade more significantly, with MOTA at only 31 and 29 at 50 Hz, respectively, and falling below 26 at 60 Hz. This result fully demonstrates that the proposed multi-dimensional data association strategy, which integrates appearance similarity, motion consistency, and cross-frame attention, effectively addresses the trajectory discontinuity and ID switching problems caused by similar target appearance and complex motion patterns in public spaces. By deeply integrating the appearance features, motion trends, and temporal association features of targets in multi-source video streams, this strategy outputs stable, continuous spatiotemporal trajectories for each individual. This is crucial for the accurate recognition of sudden health events and automatic early warning, ensuring the reliability of the entire process from target perception to event response.

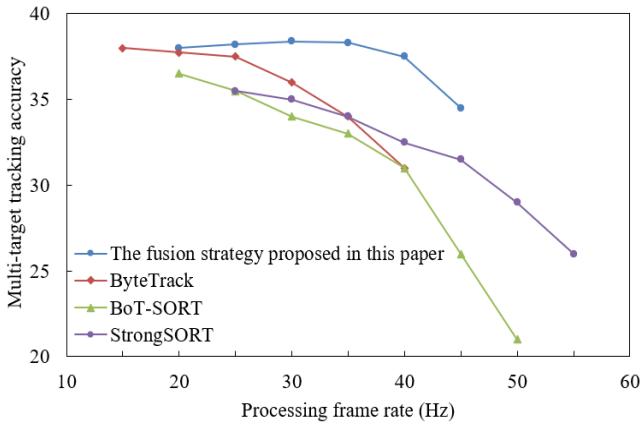

Figure 6. Relationship between multi-object tracking accuracy and processing frame rate under different data association strategies

Table 3. Multi-source video stream fusion effect experiment results

Experiment Configuration	Single-source Video Stream (View A)	Single-source Video Stream (View B)	Dual-source Video Stream Fusion (No Cross-frame Attention)	Dual-source Video Stream Fusion (With Cross-frame Attention)	Triple-source Video Stream Fusion (With Cross-frame Attention)
Trajectory Continuity (Average Number of Consecutive Tracking Frames)	45.2	42.8	68.5	75.3	82.6
ID Retention Rate (%)	68.5	65.2	82.3	88.6	92.4
Multi-view Target Matching Accuracy (%)	-	-	85.6	91.2	94.8
Target Detection Accuracy (%)	79.3	77.5	86.7	90.5	93.2
False Positive Rate (%)	12.8	14.2	8.5	5.3	3.8
Processing Frame Rate (Hz)	48.6	47.8	40.2	37.5	33.8
Performance Decay in Dense Crowd Scenarios (%)	25.3	27.1	15.2	10.1	7.5
Performance Decay in Occlusion Scenarios (%)	32.1	34.5	20.3	14.8	10.2

To verify the effectiveness of the multi-source video stream fusion strategy and the cross-frame attention module on improving target trajectory quality and recognition performance in complex scenarios, the above experiments were conducted. From the data in Table 3, it can be observed that the single-source video stream configuration performed the worst in terms of trajectory continuity, ID retention rate, and scene adaptability. The performance decay rate in dense crowd and occlusion scenarios exceeded 25%, making it insufficient to meet the full-scale monitoring needs of public spaces. With the increase in video stream number and the introduction of the cross-frame attention module, all performance metrics significantly improved: the trajectory continuity in triple-source video fusion reached 82.6 frames, ID retention rate improved to 92.4%, multi-view target matching accuracy reached 94.8%, improving by 78.3% and 34.9% compared to single-source views, while the false positive rate decreased to 3.8%. The performance decay rate in dense crowd and occlusion scenarios was controlled below 10%. Although the processing frame rate slightly decreased as the number of fused views increased, it still remained above 33.8 Hz, meeting the real-time early warning technical requirements. This demonstrates that the advantages of heterogeneous feature complementarity in multi-source video

streams, combined with the temporal feature enhancement of the cross-frame attention module, effectively solve the visibility and occlusion issues of single-source video monitoring, providing high-quality spatiotemporal data support for the accurate recognition and trajectory analysis of sudden health events.

Table 4. Automatic early warning system response performance experiment results

Experiment Configuration	Proposed System	Early Warning System without Multi-source Fusion	Early Warning System without Visualization Module	Traditional Manual Warning Simulation System
Average Warning Delay (ms)	185	160	170	3500
Warning Accuracy (%)	94.2	82.5	89.3	75.8
Fall Event Warning Coverage (%)	98.5	92.3	95.2	80.2
Syncope Event Warning Coverage (%)	96.8	88.6	93.1	75.5
Acute Cardiovascular Event Warning Coverage (%)	95.2	85.1	90.5	70.3
Occlusion Area Event Missed Detection Rate (%)	4.8	12.5	7.2	20.1
Open Area Event Missed Detection Rate (%)	2.1	5.8	3.5	15.3
Multi-event Concurrent Processing Delay (ms)	220	280	250	-

To verify the comprehensive performance of the automatic early warning system in terms of real-time, accuracy, and scene adaptability, the above comparative experiments were conducted. From the data in Table 4, it can be seen that this system outperforms the comparison systems in all core performance indicators: the average warning delay is only 185 ms, 94.7% faster than the traditional manual warning simulation system. Although slightly higher than the early warning system without multi-source fusion, the warning accuracy increased to 94.2%, a 14.3% improvement over the system without multi-source fusion. In terms of event coverage, the warning coverage for fall, syncope, and acute cardiovascular events reached 98.5%, 96.8%, and 95.2%, respectively, with missed detection rates in occlusion and open areas as low as 4.8% and 2.1%. This shows strong adaptability to different types of sudden health events and complex scenarios. In addition, the multi-event concurrent processing delay of this system is 220 ms, which is significantly lower than 280 ms for the system without multi-source fusion, meeting the emergency needs of simultaneous events in multiple areas of public spaces. The early warning system without a visualization module, due to the lack of intuitive integration of multi-source information, has weaker warning accuracy and scene adaptability than the proposed system, further proving the necessity of the "perception-recognition-warning-visualization" closed-loop framework. In conclusion, the proposed system, through the collaboration of multi-source information fusion and efficient early warning mechanisms, enables rapid and accurate early warning of sudden health events, buys critical time for emergency rescue, and significantly improves the intelligence level of public space safety management.

4. CONCLUSION

This paper addresses the technical challenges in automatic early warning systems for sudden health events in public spaces, such as dense crowd occlusion, multi-source information asynchrony, and the disconnection between early warning and decision-making. A full-process closed-loop analysis framework integrating multi-source video streams and behavior recognition algorithms is proposed. Through systematic experimental validation, this framework shows significant advantages in core performance metrics: the introduction of the cross-frame attention module improves multi-object tracking MOTA by 9.9%, reduces ID switching by 58.8%, and effectively solves the target perception robustness problem in occlusion scenarios; the behavior recognition module achieves an average recognition accuracy of 90% for fall, syncope, and other events, improving by 11.1% compared to the mainstream ST-GCN algorithm, and maintains a low missed detection rate and false positive rate in complex environments such as weak light and dense crowds; the multi-source video stream fusion strategy improves trajectory continuity by 78.3%, with ID retention rate reaching 92.4%, providing high-quality spatiotemporal data support for event recognition; the automatic early warning system achieves an average response delay of 185 ms and a warning accuracy of 94.2%, improving the efficiency of traditional manual warning systems by 94.7%. These results not only theoretically improve the technical path for multi-modal feature fusion and temporal behavior understanding in complex scenarios but also provide a practical solution with high accuracy and operability for public space safety management. The proposed system plays an important role in reducing the risk of sudden health events and enhancing urban public safety governance.

Although this framework has achieved significant progress, some limitations remain: first, the experimental data are mainly based on public datasets and simulated environments for specific scenes, and the adaptability to extreme weather and complex action interference (such as crowd playing or fall simulations) in real public spaces has not been fully verified; second, the model depends on hardware resources, and the processing frame rate drops to 33.8 Hz during triple-source video fusion, making it difficult to deploy directly on low-computing-edge devices; third, the recognition accuracy of the behavior recognition module for atypical symptoms, such as acute cardiovascular events, still has room for improvement, and it does not yet support event severity grading for early warnings. To address these shortcomings, future research can be advanced in three areas: first, constructing large-scale real-world datasets covering multiple scenes and event types,

introducing transfer learning and domain adaptation techniques to improve the model's generalization ability; second, through lightweight network design and edge computing architecture integration to reduce system deployment costs; third, integrating physiological signal sensors and video behavior features to build a multi-modal event grading recognition model to upgrade from “early warning” to “precise graded response”; fourth, exploring the application of advanced deep learning architectures such as Transformer in temporal feature extraction and multi-source information fusion to further enhance event understanding in complex scenarios.

REFERENCES

[1] Gu, C.L. (2019). Urbanization: Processes and driving forces. Science China-Earth Sciences, 62(9): 1351-1360. https://doi.org/10.1007/s11430-018-9359-y

[2] Sari, R.T., Mardiansjah, F.H. (2025). Characteristics and patterns of urbanization in non-urban regions: The case of Kudus Regency in Central Java Province, Indonesia. Journal of Regional and City Planning, 36(1): 1-21. https://doi.org/10.5614/jpwk.2025.36.1.1

[3] Solis, A.O., Wimaladasa, J., Asgary, A., Sabet, M.S., Ing, M. (2022). Shifting patterns of emergency incidents during the COVID-19 pandemic in the City of Vaughan, Canada. International Journal of Emergency Services, 11(1): 1-37. https://doi.org/10.1108/IJES-05-2021-0024

[4] Melnikova, N., Wu, J., Orr, M.F. (2015). Public health response to acute chemical incidents—Hazardous Substances Emergency Events Surveillance, nine states, 1999–2008. Morbidity and Mortality Weekly Report, 64: 25-31.

[5] Scales, S.E., Fouladi, R., Horney, J.A. (2021). Description of the use of the Incident Command System among public health agencies responding to COVID-19. Journal of Disaster Research, 16(5): 874-881. https://doi.org/10.20965/jdr.2021.p0874

[6] Hou, R., Xu, X., Dai, Y., Shao, S., Hirota, K. (2024). A multimodal fusion behaviors estimation method for public dangerous monitoring. Journal of Advanced Computational Intelligence and Intelligent Informatics, 28(3): 520-527. https://doi.org/10.20965/jaciii.2024.p0520

[7] Zhang, L.Y., Wang, M.L., Sun, H.F. (2022). Edge intelligent epidemic control system based on visual internet of things. Journal of Nonlinear and Convex Analysis, 23(9): 2049-2062.

[8] Xu, Z., Yen, N.Y., Zhang, H., Wei, X., et al. (2017). Social sensors based online attention computing of public safety events. IEEE Transactions on Emerging Topics in Computing, 5(3): 403-411. https://doi.org/10.1109/TETC.2017.2684819

[9] de Oliveira, B.V.N., de Melo, F.T. (2023). Fundamentals of computer vision: Theoretical framework of artificial recognition of images and videos. Humanidades & Inovação, 10(17): 312-327.

[10] Prijs, J., Liao, Z., Ashkani-Esfahani, S., Olczak, J., et al. (2022). Artificial intelligence and computer vision in orthopaedic trauma: The why, how, and what. The Bone & Joint Journal, 104(8): 911-914.

[11] Pervaiz, M., Shorfuzzaman, M., Alsufyani, A., Jalal, A., Alsuhibany, S.A. (2023). Tracking and analysis of pedestrian's behavior in public places. Computers, Materials & Continua, 75(1): 841-853. https://doi.org/10.32604/cmc.2023.029629

[12] Wittig, J.H., Richmond, B.J. (2014). Monkeys rely on recency of stimulus repetition when solving short-term memory tasks. Learning & Memory, 21(6): 325-333. https://doi.org/10.1101/lm.034181.113

[13] Liu, Y., Chen, L., Li, C., Liu, X., Zhou, W., Li, K. (2023). Long-term and short-term memory networks based on forgetting memristors. Soft Computing, 27(23): 18403-18418. https://doi.org/10.1007/s00500-023-09110-y

[14] Labadze, I.D., Gogoberidze, M.M., Khananashvili, M.M. (2005). Influence of partial intraspecies deprivation on short-term image-driven memory in rats. Zhurnal Vysshei Nervnoi Deiatelnosti Imeni IP Pavlova, 55(3): 368-370.

[15] Hassanpour, M., Hoseinitabatabaei, S.A., Barnaghi, P., Tafazolli, R. (2020). Improving the accuracy of the video popularity prediction models through user grouping and video popularity classification. ACM Transactions on the Web, 14(1): 4. https://doi.org/10.1145/3372499

[16] Xie, J., Zhu, Y., Chen, Z. (2021). Micro-video popularity prediction via multimodal variational information bottleneck. IEEE Transactions on Multimedia, 25: 24-37. https://doi.org/10.1109/TMM.2021.3120537

[17] Li, X., Zhou, T. (2021). Design of an online learning early warning system based on learning behaviour analysis. International Journal of Continuing Engineering Education and Life Long Learning, 31(3): 381-393. https://doi.org/10.1504/IJCEELL.2021.116035

[18] Abel, M.N., Chermak, S., Freilich, J.D. (2022). Pre-attack warning behaviors of 20 adolescent school shooters: A case study analysis. Crime & Delinquency, 68(5): 786-813. https://doi.org/10.1177/0011128721999338

[19] Zhang, F., Wang, F. (2024). Study on abnormal behaviour recognition of MOOC online English learning based on multi-dimensional data mining. International Journal of Continuing Engineering Education and Life Long Learning, 34(1): 111-122. https://doi.org/10.1504/IJCEELL.2024.135225

[20] Shen, G., Wang, J., Kong, X., Ji, Z., Zhu, B., Qiu, T. (2024). Deformation gated recurrent network for lane-level abnormal driving behavior recognition. ACM Transactions on Spatial Algorithms and Systems, 10(3): 24. https://doi.org/10.1145/3635141