Research on Elevator Entrapment Early Warning Action Recognition Method Based on YOLOv8-DA

He Li, Hongming Hu, Shengying Yang*
School of Computer Science and Technology, Zhejiang University of Science and Technology, Hangzhou 310023, China
*Corresponding email: syyang@zust.edu.cn

Abstract

Elevator entrapment incidents are frequent precursors to serious elevator accidents, posing significant risks to passenger safety. However, most existing elevator monitoring systems rely on post-event analysis and lack the ability to proactively identify abnormal passenger behaviors, especially under the confined, occluded, and visually complex conditions of elevator cabins. This poses a challenge for developing accurate and lightweight vision-based early warning methods suitable for real-time deployment. To address this issue, this paper proposes an elevator entrapment early warning action recognition method based on an improved YOLOv8 classification model, termed YOLOv8-DA. A dedicated elevator behavior dataset containing 5,501 images is constructed, covering six typical passenger behaviors related to entrapment risk. In addition, a Dual Attention Fusion Module (DAFM), integrating Efficient Channel Attention Network (ECA-Net) channel attention and Coordinate Attention mechanisms, is embedded into the YOLOv8-cls backbone to enhance both global semantic representation and fine-grained spatial discrimination. Experimental results demonstrate that YOLOv8-DA achieves an accuracy of 97.18% with only 2.7 M parameters and an inference speed of 116 FPS, outperforming representative lightweight and classical classification models. The proposed method provides an effective and practical solution for proactive elevator entrapment early warning and edge-side deployment.

Keywords

YOLOv8, YOLOv8-cls, Dual Attention Fusion Module, Efficient Channel Attention Network

Introduction

Elevators have become an indispensable component of modern urban infrastructure, serving as the primary vertical transportation system in residential buildings, commercial complexes, and public facilities worldwide. With the continuous increase in urban population density and building height, elevator usage frequency and service load have risen significantly, making elevator safety a critical concern in public transportation systems and urban management. Among various elevator-related safety incidents, elevator entrapment events occur frequently and often act as precursors to more severe accidents, posing substantial risks to passenger safety and emergency response efficiency [1].

Unlike purely mechanical failures, many elevator entrapment incidents are closely associated with passenger behaviors. In real-world scenarios, passengers often exhibit abnormal or stress-induced actions - such as forcibly prying elevator doors, excessive squatting, or leaning against car doors - before or during entrapment events. These behaviors may interfere with normal elevator operation, trigger safety protection mechanisms, or aggravate existing faults, thereby increasing the probability and severity of entrapment incidents [2]. Therefore, accurate recognition of such precursor behaviors is a key prerequisite for realizing proactive elevator safety early warning.

However, developing an effective vision-based early warning system for elevator entrapment remains challenging. Elevator cabins are confined environments with fixed camera perspectives, frequent occlusions, and complex multi-person interactions, which significantly complicate behavior recognition. In addition, lighting

variations, metallic reflections, and background interference further degrade model robustness. Moreover, the lack of publicly available datasets specifically designed for elevator entrapment precursor behaviors limits the direct application of existing behavior recognition methods trained on general-purpose datasets.

Currently, most elevator monitoring systems focus on post-event video retrieval and retrospective analysis, offering limited support for real-time risk perception and proactive intervention. Traditional rule-based or simple image-processing approaches struggle to handle complex passenger behaviors, while many deep learning-based action recognition models are designed for open or unconstrained environments and are not well suited to elevator scenarios. Although the YOLO series models have demonstrated strong performance in general visual tasks [3]. Their direct application to elevator behavior recognition is constrained by insufficient fine-grained feature discrimination and sensitivity to occlusion when used without task-specific enhancement.

To address these issues, this paper proposes a vision-based elevator entrapment early warning method based on a dual attention-enhanced YOLOv8 classification framework [4]. By combining lightweight model design with targeted channel and spatial feature enhancement, the proposed approach aims to achieve accurate and real-time recognition of entrapment precursor behaviors under complex elevator conditions, enabling proactive safety warning and efficient edge-side deployment. The remainder of this paper is organized as follows. Section 2 introduces the proposed method and the design of the Dual Attention Fusion Module (DAFM). Section 3 presents the dataset construction and experimental results, and Section 4 concludes the paper.

Theory and method

Overall framework

In this paper, we propose a novel behavior recognition model for elevator entrapment early warning - YOLOv8-DA. Considering that the core demand of elevator early warning is to judge the overall safety state of the car rather than locate individual passengers, we choose the image classification paradigm instead of the object detection paradigm: the classification model can directly output the behavior category probability of the entire image, eliminating the complex bounding box regression process, which is more suitable for the real-time deployment needs of edge devices. Based on the lightweight image classification model YOLOv8-cls, this model introduces a Dual Attention Fusion Module (DAFM) at the end of its Backbone and takes elevator car monitoring images as input [5]. This section will comprehensively elaborate on the proposed method, which consists of three key components (Figure 1).

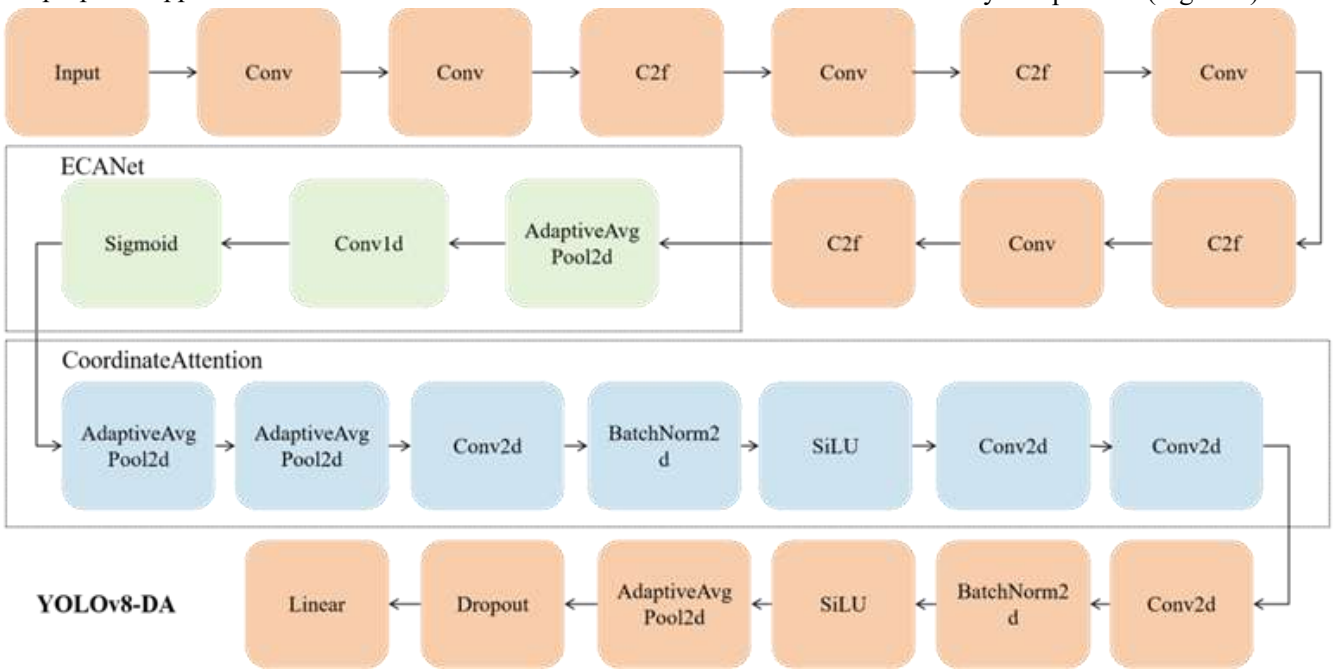

Figure 1. Structure diagram of the YOLOv8-DA model.

Figure 1 shows the structure of the proposed entrapment early warning behavior recognition framework. Firstly, an elevator car monitoring image input method is adopted, where valid images obtained by frame extraction, cleaning and manual annotation of real elevator monitoring videos are used as input, which are then fed into the proposed YOLOv8-DA hybrid model. The hybrid model is composed of three main parts: backbone feature extraction network, Dual Attention Fusion Module (DAFM), and the final Classify output classifier.

The backbone feature extraction network adopts the backbone network architecture of YOLOv8-cls, inheriting the consistent concise and efficient design concept of the YOLO series. It adopts the latest CSPNet idea and efficient cross-stage partial connections, achieving an excellent balance between computational load and feature extraction capability, and can effectively extract the basic features of passenger behaviors in elevator scenarios [6]. The Dual Attention Fusion Module (DAFM) adopts a "parallel feature weighted fusion" design: Specifically, the channel dimension generates channel attention weights through Efficient Channel Attention Network (ECA-Net), and the spatial dimension generates spatial attention weights through Coordinate Attention [7]. After the two paths of features are reduced to the same dimension by 1×1 convolution, they are weighted and summed by adaptively learned weight coefficients to realize the cooperative enhancement of global channel features and local spatial features; this dual-branch design improves the model's discriminability from both feature semantics and spatial position dimensions. In addition, this module specifically makes up for the shortcomings of the original YOLOv8-cls, such as insufficient feature extraction and sensitivity to occlusion in elevator scenarios and enhances the model's ability to capture subtle but key features in complex environments. Finally, the Classify output classifier directly outputs the probability that the entire image belongs to a certain behavior category, which perfectly meets the early warning demand of "judging the current overall safety state of the car".

The entire recognition process is divided into three stages:

(1) For the purpose of data collection and processing, surveillance videos were independently gathered from elevators of various brands and models. After frame extraction, screening, and cleaning, a specialized dataset comprising 5,501 images was constructed. The dataset covers six categories of human behavior: forcible door prying, phone use, surveillance monitoring, leaning against walls, squatting, and normal operation.

(2) Model training: We employed the Adam optimizer combined with a cross-entropy loss function. A dynamic optimization strategy was applied to key hyperparameters, including learning rate and weight initialization, to achieve stable training and optimal generalization performance.

(3) Model testing: We evaluated the performance of the trained model on the task of recognizing elevator entrapment-related behaviors, and the results validated its effectiveness and robustness.

ECA-Net channel attention

Deep convolutional neural networks have been widely applied in the field of computer vision, achieving remarkable progress in tasks such as image classification, object detection and semantic segmentation [8]. Since the groundbreaking AlexNet was proposed, continuous efforts have been made to develop new model architectures to further improve the performance of deep Convolutional Neural Networks (CNNs) [9]. In recent years, the integration of channel attention mechanisms into convolutional blocks has attracted extensive attention and demonstrated great potential for performance improvement [10]. Among these approaches, SENet is a representative method, which significantly enhances the performance of various deep CNN architectures by learning channel attention for each convolutional block, with its core operations including squeeze and excitation [11].

However, recent studies have attempted to improve the SE module by capturing more complex channel dependencies or combining additional spatial attention. Although these methods have achieved accuracy improvements, they often lead to increased model complexity and computational burden. To address this issue, Wang et al. proposed ECA-Net in 2020, which is an optimized improvement of the classic SE attention module. The core objective of ECA-Net is to effectively reduce the number of parameters and computational complexity while maintaining model performance, making it more suitable for the deployment requirements of lightweight networks.

The channel attention module is the core component of ECA-Net. Its primary function is to adaptively adjust the weight of each channel feature according to the inherent correlations between channels, thereby enhancing key features and suppressing redundant information. The workflow of this module mainly consists of three key steps:

Global average pooling: First, global average pooling is performed on the input feature map to aggregate the spatial information of each channel, yielding a channel-level feature descriptor, as shown in Equation (1):

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j) \tag{1}$$

where H and W denote the height and width of the feature map, respectively; $x_c(i,j)$ represents the pixel value of the c-th channel feature map at position (i,j); and z_c is the output result of the c-th channel after global average pooling.

Attention weight generation: The pooled channel features are transformed through a set of lightweight 1D convolutions to generate attention weights corresponding to each channel. The Sigmoid activation function is introduced to map the weight values to the range of [0,1], as shown in Equation (2):

$$\hat{z}=\sigma\left(Conv1D_k(z)\right) \tag{2}$$

where k is the kernel size of the 1D convolution; $Conv1D_k(.)$ denotes the 1D convolution operation; and z is the generated channel attention weight vector.

Feature weighting and normalization: The generated channel attention weights are applied to each channel of the original input feature map to realize the weighted combination of features from different channels. Feature normalization is completed using scaling factors to enhance the model's response to key channel features, as shown in Equation (3):

$$\tilde{x}_c=\hat{z}_c\cdot x_c \tag{3}$$

where $\hat{z}_c$ is the attention weight of the c-th channel; and $\tilde{x}_c$ is the output feature map of the c-th channel after weighting.

In the scenario of behavior recognition for elevator entrapment early warning, behaviors such as "wall leaning" usually lead to systematic changes in specific appearance features of the entire surveillance frame, which are characterized by globality and integrity. By analyzing the aggregated information of all channels, ECA-Net can enhance the model's ability to perceive such changes in global feature patterns, making the model more sensitive to the behavioral representations that rely on overall scene features, thereby improving the recognition accuracy of such behaviors.

Coordinate attention spatial attention

Currently, the SE module proposed by SENet remains the most widely used attention mechanism in lightweight networks. By calculating channel attention via 2D global pooling, it achieves significant performance improvements at a low computational cost. However, the SE module has obvious limitations: It only focuses on information encoding between channels and completely ignores the critical role of position information, which is essential for visual tasks that require accurate capture of target structures and local details. Subsequent methods such as CBAM have attempted to introduce position information by reducing the number of channels and applying large-size convolutions, but convolution operations can only capture local correlations, leading to unsatisfactory performance in modeling long-range dependencies, which are core to visual tasks [12].

To solve this problem, this paper introduces an efficient novel attention mechanism, namely Coordinate Attention. Its core design concept is to integrate position information into channel attention, enabling lightweight networks to perform attention operations over a wider range of regions while avoiding excessive computational overhead. To alleviate the loss of position information caused by 2D global pooling, Coordinate Attention decomposes channel attention into two parallel 1D feature encoding processes, effectively incorporating spatial coordinate information into the generated attention maps. Its specific workflow is as follows:

1D global pooling: Two independent 1D global pooling operations are adopted to aggregate information from the input feature map along the vertical height and horizontal width directions, respectively, forming two direction-aware feature maps that retain spatial coordinate information, as shown in Equation (4) and (5):

$$z_c^{\mathrm{h}}(\mathrm{h})=\frac{1}{W}\sum_{0\le j<W}x_c(\mathrm{h},j) \tag{4}$$

$$z_c^{w}(w)=\frac{1}{H}\sum_{0\le i<H}x_c(i,w) \tag{5}$$

where $z_c^{\mathrm{h}}(w)$ denotes the pooling result of the c-th channel at position h along the vertical direction; $z_c^{w}(w)$ denotes the pooling result of the c-th channel at position

w along the horizontal direction; and H and W are the height and width of the feature map, respectively.

Coordinate information transformation and fusion: The two feature maps embedded with specific directional information are concatenated, and feature transformation is performed via 1D convolution. The ReLU activation function is introduced to enhance nonlinear expression capability, yielding the fused feature vector, as shown in Equation (6):

$$\boldsymbol{f}=\delta(F_1([\mathbf{z}^h,\mathbf{z}^w])) \quad (6)$$

where $[\mathbf{z}^h, \mathbf{z}^w]$ represents the concatenation operation of the vertical feature map $\mathbf{z}^h$ and the horizontal feature map $\mathbf{z}^w$; F_1 denotes the 1D convolution transformation; and f is the feature vector fused with coordinate information.

Attention map encoding: The fused feature vector is split into two independent feature branches corresponding to the vertical and horizontal directions, respectively. Two attention maps are generated via encoding using respective 1D convolutions (F_h ,F_w) and the Sigmoid activation function σ, as shown in Equation (7) and (8):

$$\mathbf{g}^h=\sigma(F_h(\mathbf{f}^h)) \quad (7)$$

$$\mathbf{g}^w=\sigma(F_w(\mathbf{f}^w)) \quad (8)$$

where f^h and f^w are the vertical and horizontal directional branches obtained by splitting the feature vector f, respectively; g^h and g^w are the vertically and horizontally oriented attention maps generated by encoding, respectively; each attention map captures long-range dependencies along the corresponding spatial dimension in the input feature map.

Feature enhancement: The two generated attention maps are synchronously applied to the original input feature map, and feature enhancement is realized via pixel-wise multiplication, as shown in Equation (9):

$$\tilde{x}_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \quad (9)$$

where $g_c^h(i)$ is the weight value of the vertically oriented attention map of the c-th channel at position i; $g_c^w(j)$ is the weight value of the horizontally oriented attention map of the c-th channel at position j; and $\tilde{x}_c(i,j)$ is the pixel value of the enhanced feature map at position (i, j). Compared with previous attention mechanisms applied to lightweight networks, Coordinate Attention has significant advantages: (1) It simultaneously possesses the capabilities of cross-channel information capture, direction awareness and position awareness, which can improve both the target localization accuracy and recognition accuracy of the model. (2) It features a flexible and lightweight structural design, which can be seamlessly integrated into classic lightweight architecture such as the inverted residual module of MobileNetV2 and the hourglass module of MobileNeXt, effectively enhancing feature representation capability [13]. (3) It can bring significant performance improvements to downstream tasks of lightweight networks, especially showing outstanding performance in dense prediction tasks such as semantic segmentation.

In the elevator entrapment early warning scenario, the discriminative features of key behaviors such as “door prying” and “phone calling” are often concentrated in very specific and small spatial regions, which are easily interfered by complex backgrounds or occlusions. By aggregating features along the X-axis and Y-axis coordinates separately, Coordinate Attention can accurately capture the correlations between pixels far away from the current position, thereby penetrating interference information and directly focusing attention on these key local regions. This is crucial for locating small-target behaviors in complex car environments and improving the recognition accuracy of such high-risk behaviors.

Experiment

Experimental setup

All experiments were conducted under the same hardware and software environment to ensure fairness and reproducibility. The experimental platform consisted of a 14th-generation Intel Core i9-14900HX processor, an NVIDIA GeForce GTX 4060 GPU, Windows 11 operating system, CUDA 12.4, and the PyTorch 1.13.0 framework.

The models were trained using the Adam optimizer with the cross-entropy loss function. Classification accuracy for each behavior category, as well as overall accuracy, were adopted as evaluation metrics. The main training hyperparameters of the proposed YOLOv8-DA model are summarized in Table 1, including a learning rate of 0.001, 200 training epochs, and a batch size of 32. All models were trained under the same experimental settings to ensure a fair performance comparison.

Table 1. Hyperparameter settings of YOLOv8-DA.

Hyperparameters	Value
Learning rate	0.001
Epochs	200
Batch size	32

Dataset

When testing the proposed model, self-collected and constructed image data were adopted. This dataset was obtained via frame extraction from video footage captured by surveillance cameras installed on elevators of various brands and models, with the surveillance equipment illustrated in Figure 2. A total of 857 segments of elevator entrapment videos were collected, with each video lasting approximately ten minutes.

Figure2. Monitoring equipment.

After screening and data cleaning, 5,501 images of passenger entrapment behaviors were finally obtained. Based on the video content, passenger behaviors were categorized into six classes, namely door prying, phone calling, surveillance camera watching, wall leaning, squatting, and normal behavior. The dataset of elevator passenger entrapment behaviors is detailed in Table 2.

Table 2. Elevator passenger entrapment behavior dataset.

Data type	Dataset size
Door prying	1,050
Phone call	436
Monitoring	406
Wall hugging	1,840
Squat	1,022
Normal	747

When evaluating the performance of the YOLOv8-DA model, selecting appropriate comparison models is crucial. In this paper, YOLOv8-cls, AlexNet, VGG, and LeNet-5 are selected as comparison models, covering the network spectrum from shallow to deep and from simple to complex, which can comprehensively evaluate the capability of YOLOv8-DA [13]. These models represent different design approaches respectively: AlexNet and VGG emphasize depth and structural simplicity, while LeNet-5 reflects a lightweight but limited-capability design. This enables a comprehensive evaluation of the comprehensive advantages of YOLOv8-DA in terms of accuracy, efficiency, parameter quantity, and other aspects. Detailed comparative experimental data are shown in Table 3.

Table 3. Comparative experiments on the elevator passenger entrapment behavior dataset.

Model	Parameters	Inference time	FPS	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
AlexNet	57.0 M	20.0 ms	49.50	96.72	96.73	96.70	96.73
VGG	128.8 M	17.4 ms	57.40	94.70	94.76	94.71	94.76
LeNet-5	61.7 K	1.0 ms	997.56	91.10	91.26	91.15	91.26
YOLOv8-cls	2.7 M	8.4 ms	118.20	96.19	96.18	96.18	96.18
YOLOv8-DA	2.7 M	8.6 ms	116.40	97.20	97.18	97.18	97.18

The comparison of the parameter quantities of each model, as well as the convergence patterns and classification performances of different evaluation models that show significant differences. The proposed YOLOv8-DA model exhibits instability in the initial stage but becomes stable in the later stage of training. While maintaining the second smallest parameter quantity, it achieves the optimal performance of 97.18%, indicating that the method maintains the lightweight nature of the model while achieving the highest accuracy. Compared with the model in this chapter, YOLOv8-cls has poorer stability, with a final accuracy of 96.18%.

AlexNet quickly achieves high accuracy of 96.73% in the early stage but fails to further improve in the later stage of training, and its parameter quantity is the second largest among the comparison models. VGG has an extremely large parameter quantity, far exceeding other models, while its accuracy only reaches 94.76%.

Finally, LeNet-5, with the smallest parameter quantity, also has the lowest model accuracy of 91.26%. To strictly evaluate the classification performance of the models on complex samples, this study adopted a confusion matrix for identification analysis (as shown in Figure 3-7).

Several other comparison models lack sufficient ability to judge the two behaviors of normal behavior and watching surveillance, but the YOLOv8-DA model proposed in this paper can greatly improve the accuracy of these two behaviors, reaching 93.67% and 95.51% respectively.

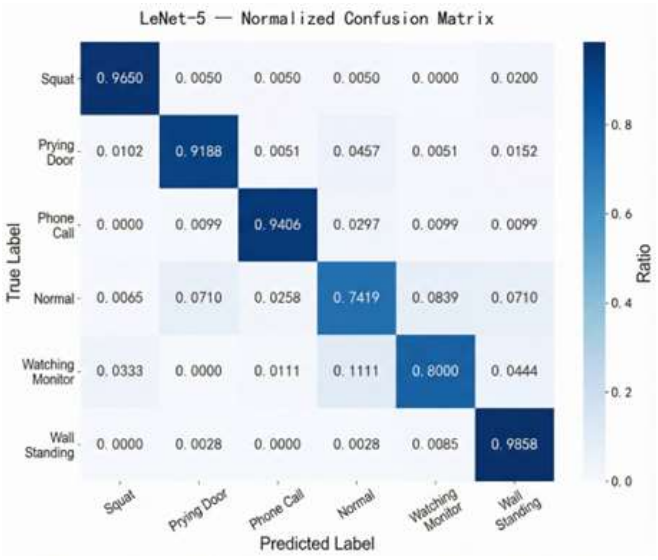

Figure3. Confusion Matrix of LeNet-5.

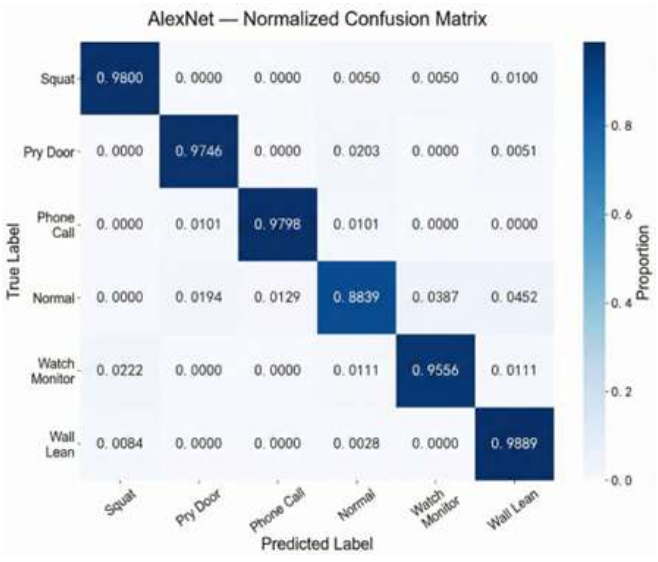

Figure 4. Confusion Matrix of AlexNet.

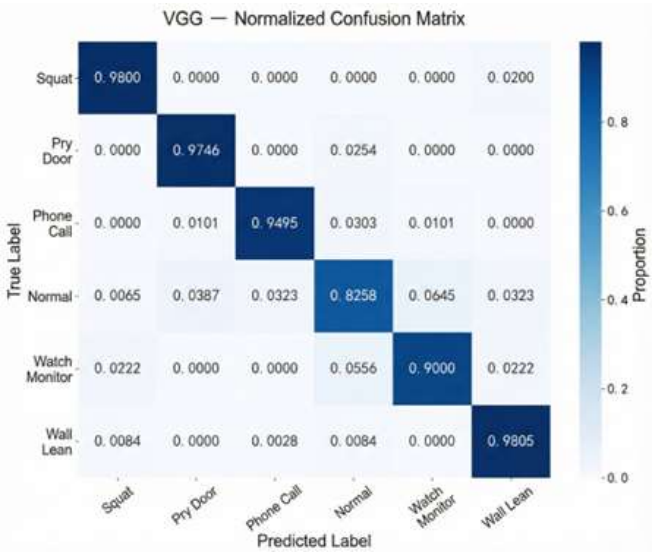

Figure 5. Confusion Matrix of VGG.

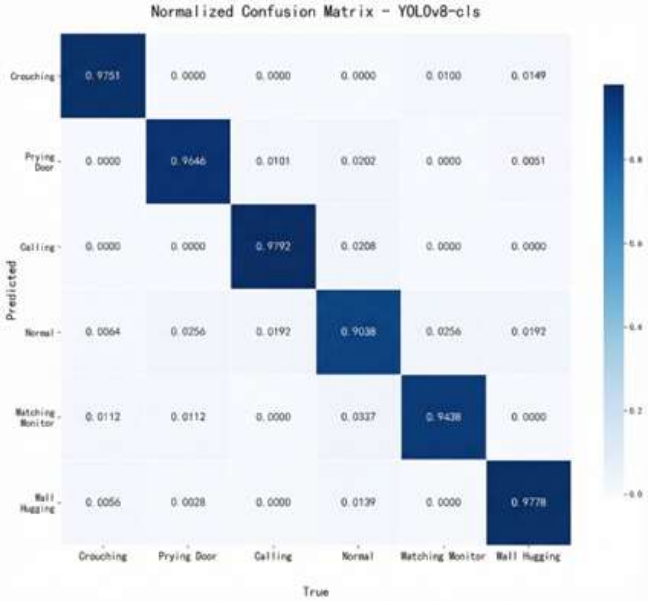

Figure 6. Confusion Matrix of YOLOv8-cls.

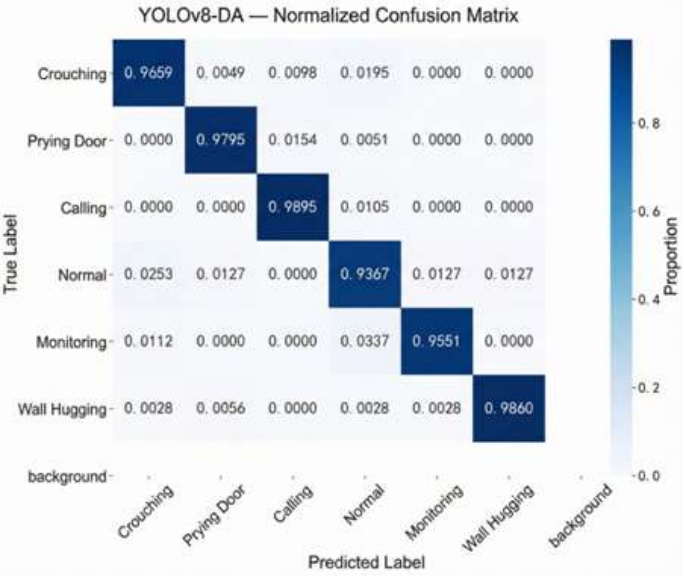

Figure.7. Confusion Matrix of YOLOv8-DA

To further explore the decision-making mechanism and

feature focusing patterns of different deep learning models in the task of elevator entrapment behavior recognition, this study adopted Grad-CAM technology to conduct an in-depth visual analysis on each comparison model [14]. By comparing the activation heatmaps of four typical behaviors, the essential differences in feature perception and semantic understanding among different model architectures were revealed (as shown in Table 4).

Table 4. Results of ablation experiments.

Method configuration		Detection accuracy (%)						
ECA	CA	Overall	Squat	Door prying	Phone call	Normal	Monitoring	Wall hugging
/	/	96.18	97.51	96.46	97.92	90.38	94.38	97.78
/	√	96.45	97.52	95.57	98.96	93.92	95.51	97.51
√	√	96.09	95.61	96.45	97.92	89.81	95.45	98.60
√	/	97.18	96.59	97.95	98.95	93.67	95.51	98.60

The proposed YOLOv8-DA model demonstrates superior attention distribution compared with the baseline and classical models. The Grad-CAM visualizations show that YOLOv8-DA can accurately focus on key body parts related to passenger behaviors while simultaneously capturing relevant elevator scene context. This indicates that the introduced dual attention mechanism effectively achieves collaborative optimization in the channel and spatial dimensions, enabling the model to perceive both local discriminative details and global scene semantics, thereby improving recognition accuracy.

In contrast, AlexNet tends to focus excessively on elevator structural elements such as doors and cabin corners, showing limited attention to human behavioral regions, which reflects its relatively shallow feature extraction capability. LeNet-5 exhibits highly diffused attention across the background due to its limited receptive field, failing to localize behavior-related regions effectively. VGG shows moderate performance by attending to both human and environmental features, but its attention remains scattered and lacks the precise behavior-scene association observed in YOLOv8-DA. The baseline YOLOv8-cls model presents balanced global attention to both human bodies and elevator environments. However, without dedicated attention enhancement, it primarily captures coarse semantic information and struggles to focus on fine-grained local interactions, such as finger-door contact in door-prying behaviors (as shown in Figure 8-10).

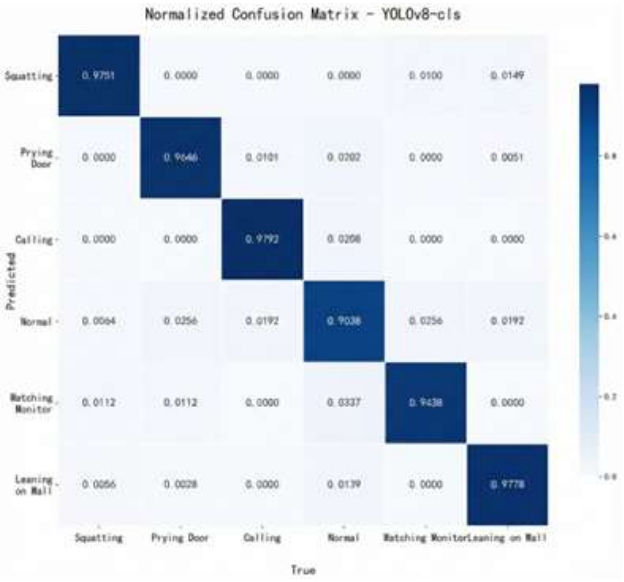

Figure 8. Confusion matrix of YOLOv8-cls.

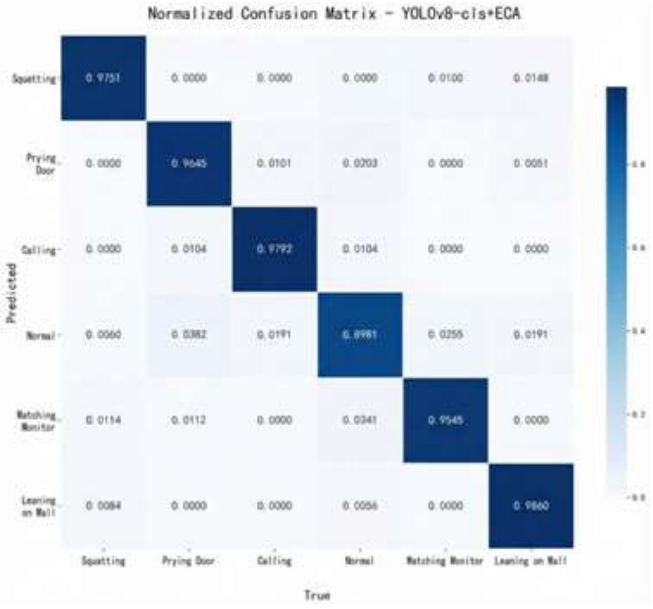

Figure 9. Confusion matrix of YOLOv8-cls+ECA.

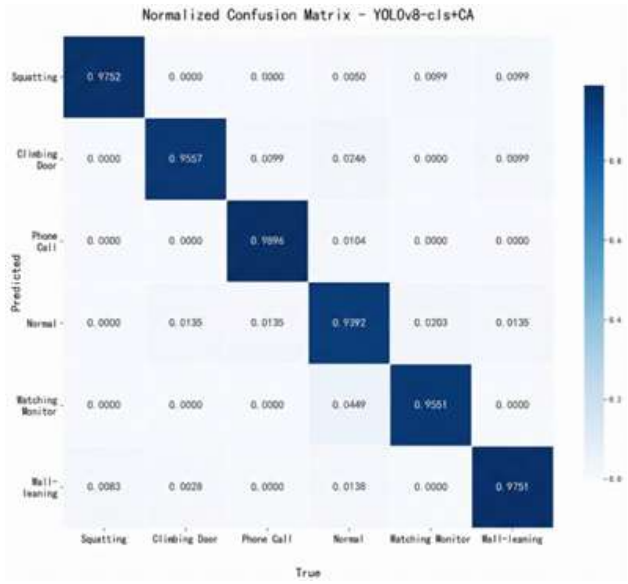

Fig.10. Confusion matrix of YOLOv8-cls+CA.

Overall, classical models and the YOLOv8-cls baseline suffer from either overly global or scattered attention distributions. By contrast, YOLOv8-DA achieves more focused and discriminative attention through its dual attention mechanism, providing an intuitive explanation for its superior performance.

Ablation experiment

To further analyze the contribution of different attention components, ablation experiments were conducted by selectively removing the ECA and Coordinate Attention (CA) modules. As shown in Table 4, the complete YOLOv8-DA model achieves the highest overall accuracy. Removing the CA module leads to a 1.08% drop in accuracy, while removing the ECA module results in a 0.73% decrease, indicating that both modules contribute positively to model performance.

Specifically, the CA module significantly improves the recognition of behaviors requiring precise spatial localization, such as phone calling and normal behaviors, while the ECA module mainly enhances behaviors dominated by global appearance features, such as wall leaning. When combined, the dual attention mechanism yields consistent performance gains across almost all behavior categories, demonstrating the complementary roles of channel and spatial attention in elevator entrapment behavior recognition (as shown in Figure 11-13).

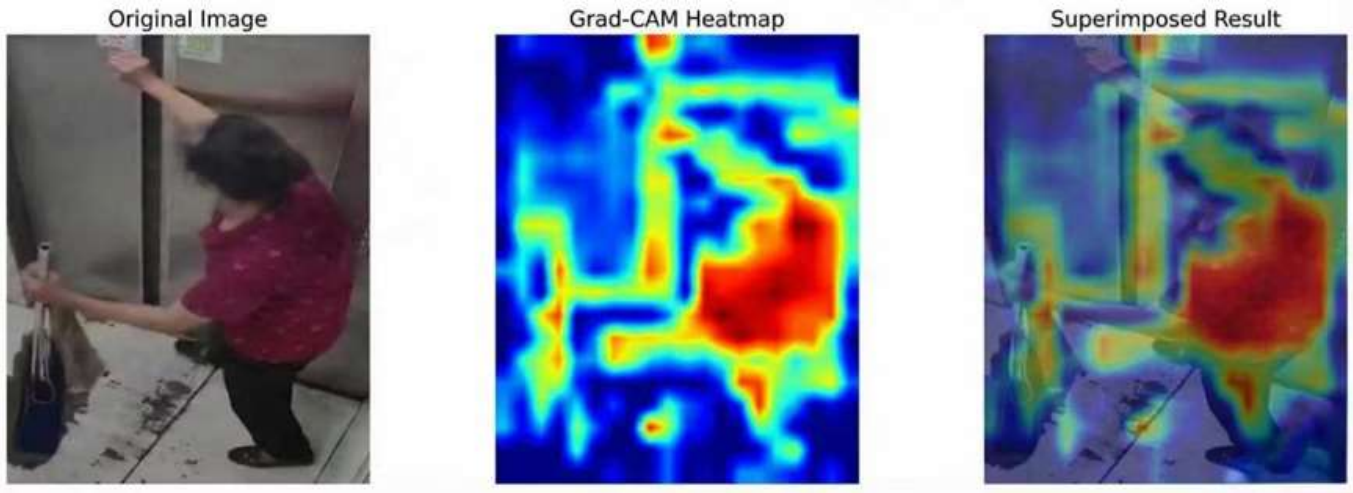

Figure11. Grad-CAM Visualization Diagram of YOLOv8-cls.

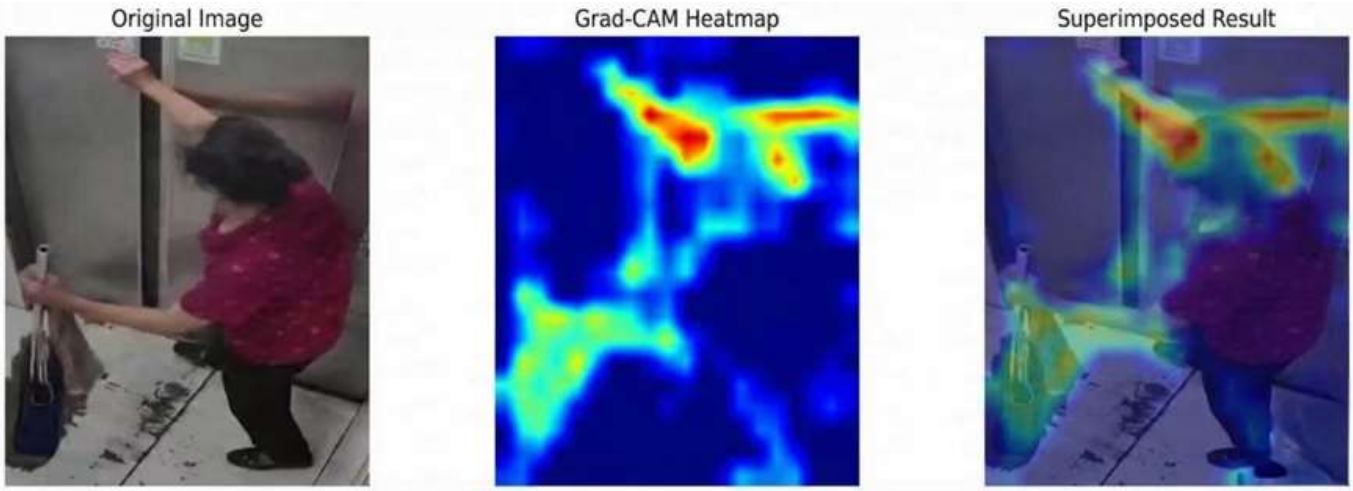

Figure12. Grad-CAM Visualization Diagram of YOLOv8-cls+ECA.

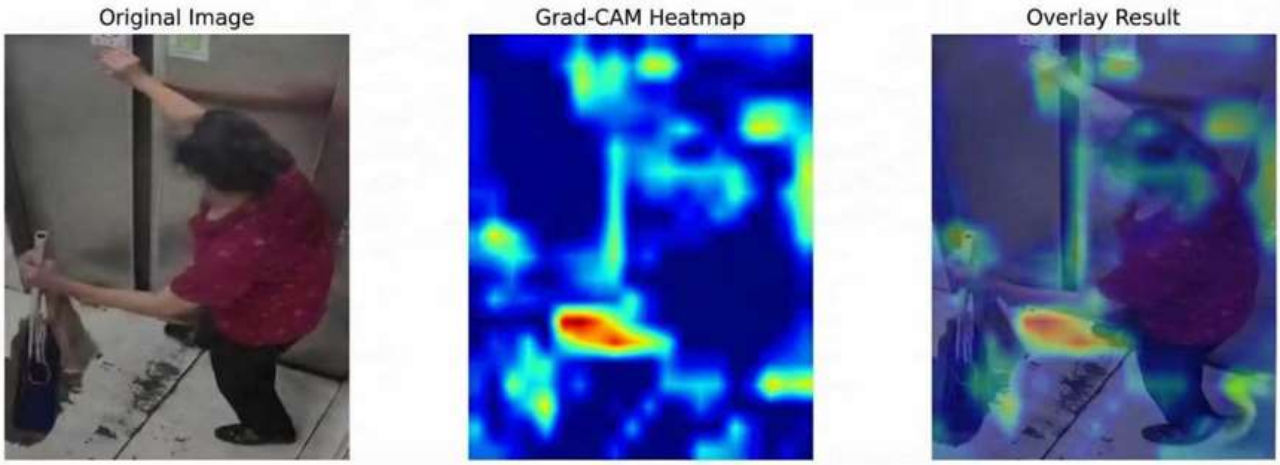

Figure 13. Grad-CAM Visualization Diagram of YOLOv8-cls+CA.

As the baseline, YOLOv8-cls focuses mainly on global scene elements, such as human silhouettes and elevator structures, but fails to capture fine-grained local interactions like hand–door contacts, limiting its ability to distinguish similar behaviors.

Adding the ECA-Net module enhances channel-wise attention, improving focus on key body parts, such as upper limbs, and better representing overall behavioral patterns. Introducing Coordinate Attention further refines spatial localization, highlighting small but discriminative regions like fingertip-door contacts and enabling the model to handle occlusions and complex backgrounds effectively. This coarse-to-fine attention evolution demonstrates the complementary roles of channel and spatial mechanisms in capturing both global semantics and local details

Conclusion

This study proposes a dedicated behavior recognition method for elevator entrapment early warning, featuring a self-constructed dataset of 5,501 images and a lightweight YOLOv8-DA model with a dual attention mechanism. The method addresses key challenges in elevator scenarios, including data scarcity, severe perspective interference, and complex environments, enabling accurate and real-time recognition of precursory behaviors.

YOLOv8-DA enhances the YOLOv8-cls backbone with a Dual Attention Fusion Module (DAFM), combining ECA-Net for global channel attention and Coordinate Attention for local spatial focus. This design achieves efficient feature extraction and classification while maintaining a lightweight 2.7 M-parameter model, suitable for edge deployment.

Experimental results demonstrate an overall accuracy of 97.18%, outperforming YOLOv8-cls (96.18%) and classical models such as AlexNet (96.73%) and VGG (94.76%). The model excels in distinguishing challenging behaviors like normal and surveillance-watching actions, with accuracies of 93.67% and 95.51%, respectively. Ablation studies confirm the complementary contributions of the ECA and CA modules, and Grad-CAM visualizations verify precise attention to key behavioral parts and scene context.

Field validation over six months in three residential communities shows a ≤10 s early warning response, a 40% reduction in entrapment incidents, and a 55% decrease in secondary panic behaviors. This demonstrates the method's practical value in enhancing elevator safety, shifting from post-hoc monitoring to proactive risk detection and early warning in real-world deployments.

Funding

This work was not supported by any funds.

Acknowledgements

The authors would like to show sincere thanks to those techniques who have contributed to this research.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Prahlow, J. A., Ashraf, Z., Plaza, N., Rogers, C., Ferreira, P., Fowler, D. R., Lantz, P. E. (2020) Elevator-related deaths. *Journal of Forensic Sciences*, 65(3), 823-832.

[2] Zhenbo, C., Mu, Y., Haoxin, Z., Xuesong, X., Wenchao, L., Kun, S., Gang, X. (2025) Review of

fault modes, diagnosis and prediction methods for elevator systems. *Journal of Vibration Engineering & Technologies*, 13(7), 473.

[3] Ali, M. L., Zhang, Z. (2024) The YOLO framework: a comprehensive review of evolution, applications, and benchmarks in object detection. *Computers*, 13(12), 336.

[4] Zhang, C., Li, Z., Li, J., Zou, L., Dong, E. (2025) Optimization of visual detection algorithms for elevator landing door safety-keeper bolts. *Machines*, 13(9), 790.

[5] Luo, J., Yang, X., Dai, Q., Qiu, W., Nie, S., Wu, J., Zeng, M. (2025) Multimodal fusion-based self-calibration method for elevator weighing towards intelligent premature warning. *Sensors*, 25(17), 5550.

[6] Wang, Z., Chen, J., Yu, P., Feng, B., Feng, D. (2024) SC-YOLOv8 network with soft-pooling and attention for elevator passenger detection. *Applied Sciences*, 14(8), 3321.

[7] Liu, H., Zhang, Y., Chen, Y. (2024) A symmetric efficient spatial and channel attention (ESCA) module based on convolutional neural networks. *Symmetry*, 16(8), 952.

[8] Guo, Y., Liu, Y., Georgiou, T., Lew, M. S. (2018) A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(2), 87-93.

[9] Khan, A., Sohail, A., Zahoora, U., Qureshi, A. S. (2020) A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, 53(8), 5455-5516.

[10] Zhang, S., Liu, Z., Chen, Y., Jin, Y., Bai, G. (2023) Selective kernel convolution deep residual network based on channel-spatial attention mechanism and feature fusion for mechanical fault diagnosis. *ISA Transactions*, 133, 369-383.

[11] Altuwaijri, G. A., Muhammad, G., Altaheri, H., Alsulaiman, M. (2022) A multi-branch convolutional neural network with squeeze-and-excitation attention blocks for EEG-based motor imagery signals classification. *Diagnostics*, 12(4), 995.

[12] Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J. (2021) Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems*, 34, 30008-30022.

[13] Hassan, H. Z., Saeed, N. M. (2024) Advancements and applications of lightweight structures: a comprehensive review. *Discover Civil Engineering*, 1(1), 47.

[14] Nguyen, H. (2020) Fast object detection framework based on mobilenetv2 architecture and enhanced feature pyramid. *J. Theor. Appl. Inf. Technol*, 98(05), 812-824.