RESEARCH ARTICLE

An Enhanced YOLOv9-Based Detection Method and Warning System for Indoor Electric Motorcycles

Qin Luo[1], Pengxiang Lu[1], Yibo Sun[1], Zhichao Yin[2] and Moufa Hu[2,*]

[1] College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China
[2] Key Laboratory of Automatic Target Recognition, National University of Defense Technology, Changsha 410073, China

Abstract

To address severe fire safety risks caused by electric motorcycles (EMs) and their batteries being illegally brought into building elevators, this paper presents a real-time EM detection and alarm system for elevator environments, built upon a multi-source information fusion framework and an improved YOLOv9. To elevate detection accuracy for EMs in confined elevator spaces, two core optimizations are embedded into the network: the Programmable Gradient Information (PGI) training strategy, and a lightweight Generalized Efficient Layer Aggregation Network (GELAN) backbone enhanced with depthwise separable convolution (DSConv). A dedicated dataset consisting of roughly 2,000 images is established for model training and validation. This dataset covers a wide range of elevator scenes, diverse target postures, and common occlusion cases. Experimental results show that the proposed model achieves mAP@0.5 values of 0.952, 0.915 and 0.887 across three challenging elevator scenarios. The average per-frame inference latency is less than 35 ms, which fully meets the real-time constraints for edge device deployment. To construct a complete closed-loop detection and alarm system, we further design an alarm subsystem empowered by multi-source information fusion. It adopts dual-threshold decision logic and supports multimodal alerts including audio, visual cues, and log recording. This framework first judges the presence of EMs via a basic confidence threshold, then verifies the temporal continuity of detection results using a dedicated alarm threshold, and finally triggers multimodal alarms accordingly. This system effectively improves the efficiency of alarm notifications and provides a practical technical solution for fire safety management of elevators in residential and commercial buildings.

Keywords: electric motorcycles detection, YOLOv9, real-time system, intelligent alarm, multi-source information fusion.

Academic Editor:
Guoyan Wang

Submitted: 15 January 2026
Accepted: 08 June 2026
Published: 22 June 2026

Vol. 3, **No.** 2, 2026.
10.62762/CJIF.2026.420972

***Corresponding author:**
Moufa Hu
mfhu20080630@nudt.edu.cn

1 Introduction

With the acceleration of urbanization and changes in residents' travel patterns, electric motorcycles (EMs) have become increasingly prevalent due to

Citation
Luo, Q., Lu, P., Sun, Y., Yin, Z., & Hu, M. (2026). An Enhanced YOLOv9-Based Detection Method and Warning System for Indoor Electric Motorcycles. *Chinese Journal of Information Fusion*, 3(2), 153–165.

their convenience, affordability, and environmental friendliness, leading to a dramatic surge in their market ownership, which has in turn motivated the development of deep learning-based deterrence systems to restrict their unauthorized entry into shared indoor spaces such as building entrances and elevators [1]. However, this rapid growth has brought significant challenges to community safety management, particularly regarding the illegal transportation of EMs into residential building elevators and subsequent indoor charging in corridors or private residences. Statistics show that a substantial proportion of annual fire incidents in China are EM-related [2], with over 80% of these fires occurring during the charging process, primarily triggered by battery thermal runaway. Frequent spontaneous combustion incidents have made the practice of bringing EMs "upstairs" or into residential units a major fire safety hazard for residents [3]. Such fires are characterized by rapid flame spread, massive toxic fume emission, and a high risk of casualties and property damage, posing severe challenges to community governance and emergency management.

In response to this critical safety hazard, national and local authorities in China have issued a series of regulations and normative documents. For instance, the Several Regulations on Fire Safety Management in Urban Residential Areas of Hunan Province explicitly prohibits indoor EM charging and the carriage of EMs or their batteries into elevators. Nevertheless, the enforcement of these regulations relies heavily on manual supervision and resident compliance, lacking efficient, real-time, and automated technical deterrence measures [4]. Currently, most methods for preventing EM entry into elevator cabins rely on image recognition technology. However, for complex electromechanical environments such as elevators, relying solely on single-source data from visual sensors for decision-making in intricate scenarios is unreliable and insufficient to guarantee adequate detection accuracy [5]. Traditional community security systems (e.g., standard surveillance cameras) typically only provide video recording and post-incident traceability functions, and are unable to issue immediate warnings or implement intervention measures when violations occur. Therefore, developing an intelligent detection and alarm system that can automatically, accurately, and in real time identify illegal EM entry into elevators, and trigger on-site warnings and linked control actions, has urgent practical demand and significant engineering application value.

The main contributions of this paper are summarized as follows: Custom Elevator Dataset: A specialized dataset is constructed for EM detection in elevator environments. It covers diverse lighting conditions (bright and dim), target states (complete targets, human-vehicle interaction, and partial occlusion), supporting robust model training and comprehensive performance evaluation.

Enhanced YOLOv9 algorithm: We adopt the Programmable Gradient Information (PGI) training strategy to mitigate information loss during forward propagation, and integrate depthwise separable convolution (DSConv) into the GELAN backbone of YOLOv9 to realize model lightweighting. DSConv splits standard convolution into depthwise convolution for spatial feature extraction and pointwise convolution for channel feature fusion. Its total computational cost is reduced to 1/N+1/9 of standard convolution. The optimized model achieves a per-frame inference time below 35 ms across various test scenarios.

Multi-source Information Fusion Alarm System: A multi-source information fusion framework is designed for real-time EM detection and alarming. The system consists of a dual-threshold decision module that fuses detection confidence and temporal continuity, and a multimodal alarm module that synchronizes audio alerts, visual warnings, and event logging. This integrated framework forms a reliable closed-loop pipeline and greatly improves the stability of elevator fire safety monitoring.

The remainder of this paper is organized as follows: Section 2 reviews relevant technologies and existing research on EMs detection in elevator scenarios. Section 3 details the overall system architecture, core algorithm principles, and optimization strategies. Section 4 describes the experimental design and result analysis, including dataset construction, model training, and performance evaluation. Section 5 concludes the paper and discusses future research directions.

2 Related Work

As a mainstream branch of computer vision, deep learning-based object detection provides a feasible technical approach for intelligent monitoring [6]. The YOLO series formulates object detection as a single-stage regression task, achieving an excellent trade-off between detection accuracy and inference speed [7]. Benefiting from this advantage, YOLO algorithms are widely applied in real-time security surveillance. From YOLOv1 to YOLOv9, continuous improvements have been made to backbone networks, feature fusion modules, and loss functions, leading to steady performance gains [8].

Nevertheless, directly applying a generic object detection model to the specific vertical scenario of building elevators still faces several non-negligible challenges [9]. First, the elevator cabin is a confined space with limited camera viewing angles; EMs may appear in various poses and scales, and are often severely occluded by passengers or luggage, resulting in incomplete target feature information. Second, lighting conditions inside elevators are complex and variable, ranging from bright daylight to nighttime illumination solely by ceiling lights, and even include overexposure or underexposure caused by direct light or backlighting, which severely degrades image quality and detection stability [10, 11]. Third, EMs and their batteries have a wide variety of appearances, and battery targets are relatively small, further increasing the difficulty of accurate detection. Finally, the detection system must meet stringent real-time requirements (typically completing the full detection-alarm loop within hundreds of milliseconds) and operate stably on resource-constrained edge computing devices.

To address these challenges, recent studies have explored the application of object detection algorithms to the problem of illegal EM entry into elevators [12]. For example, to solve the problems of the YOLOv4 model being too large for embedded deployment and having slow recognition speed, Liu et al. [13] replaced the backbone network of YOLOv4 with MobileNetv2. Lin et al. [14] proposed an online EM detection system for elevator corridors based on an improved YOLOv4-Tiny algorithm to reduce misdetection, such as mistaking mountain bikes with thick tires for EMs. Wan et al. [15] optimized the loss function of YOLOv8 by adopting the Inner-CIoU loss function to enhance the model's detection performance and generalization ability. Other studies have implemented EM detection and tracking based on YOLOv5 or improved DeepSORT algorithms [16, 17]. Wang et al. [18] proposed the concept of Programmable Gradient Information (PGI) to address the various modifications required for deep networks to achieve multiple objectives. PGI provides complete input information for the target task, which is used to compute the objective function, thereby obtaining reliable gradient information for updating the network weights. Additionally, a novel lightweight network architecture based on gradient path planning was designed. Compared to YOLO MS, YOLOv7 AF, and YOLOv8-X, YOLOv9 reduces parameters by approximately 10%, 42%, and 16%, respectively, while achieving AP improvements of 0.4–0.6%, no change, and 1.7%. Ikmel and El Amrani [19] compared the effectiveness of YOLOv5, YOLOv7, YOLOv8, and YOLOv9 in object detection within road environments. The results indicate a gradual improvement in object detection performance from YOLOv5 to YOLOv9. YOLOv9 achieves the shortest inference time, outperforming previous versions while maintaining competitive accuracy and recall. Similarly, KIM et al. [24] investigated a YOLOv5-based real-time object detection system for elevator monitoring, further confirming the practical feasibility of deep learning-driven elevator safety solutions.

Nevertheless, most existing methods are validated on general datasets or under ideal laboratory conditions. Few studies focus on model robustness against elevator-specific interferences such as complex lighting and heavy occlusion. Moreover, most current research stays at the algorithm level, lacking complete engineering implementation and verification of integrated alarm systems. Designing intelligent alarm logic to suppress false alarms also remains an open problem. Targeting the above gaps, this paper develops a full-stack detection and alarm solution based on enhanced YOLOv9 for elevator scenarios.

3 Methodology

3.1 Multi-Source Information Fusion Framework Based on Enhanced YOLOv9

The proposed system adopts a hierarchical and modular architecture centered on multi-source information fusion, tailored for real-time EM intrusion detection in confined elevator environments. As illustrated in Figure 1, the overall system is divided into three closely coupled layers, which cooperate to form a complete closed-loop safety management pipeline: the Perception Layer acquires real-time visual data, the Decision Layer runs detection algorithms and fusion decision logic, and the Execution Layer outputs multimodal warnings and records event logs.

The Perception Layer, responsible for continuous real-time visual data acquisition, employs industrial high-definition wide-angle cameras mounted on the elevator ceiling. The cameras support 1080p resolution and 30 frames per second (FPS), providing a full cabin

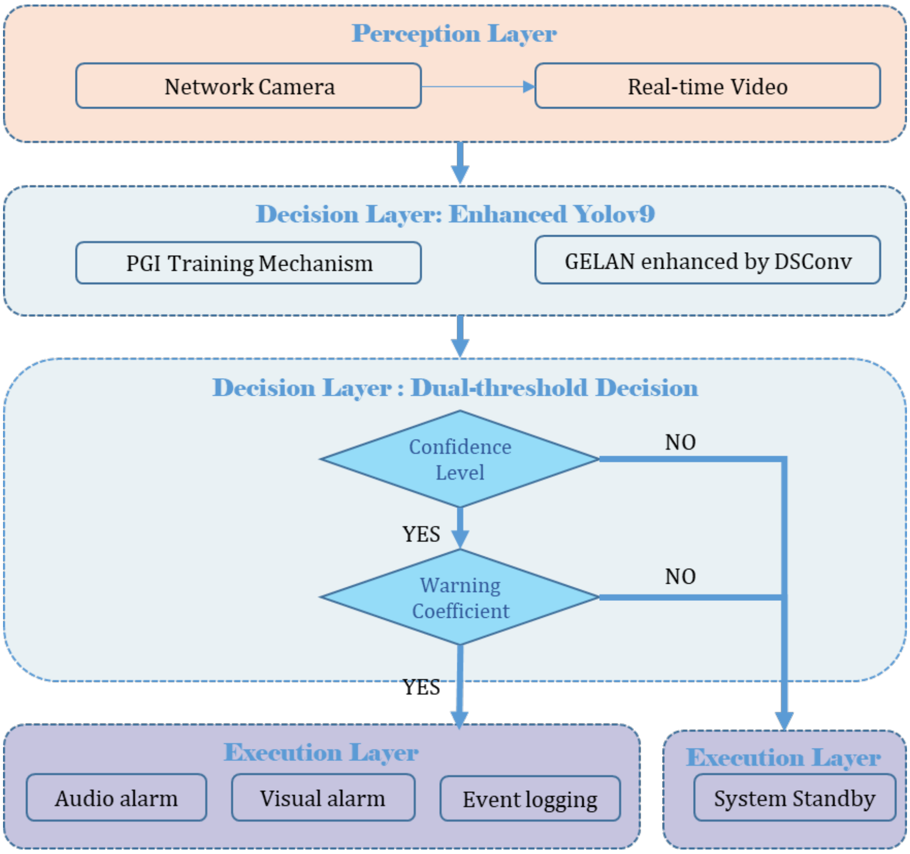


Figure 1. Overall architecture and workflow of the EMs detection and alarm system.

field of view with minimal distortion. Captured video streams are transmitted to the Decision Layer through a low-latency local network for subsequent analysis.

The Decision Layer constitutes the core of the entire system, undertaking real-time video stream analysis and intelligent alarm judgment. It integrates two key components: an enhanced YOLOv9 object detection model optimized for elevator scenarios and a dual-threshold multi-source information fusion module. We replace standard convolutions in the GELAN backbone with DSConv to realize model lightweight while preserving feature extraction capability. The computational cost is reduced to 1/N+1/9 of conventional convolution. The dual-threshold multi-source information fusion decision module operates as follows: first, a confidence threshold is applied to filter out low-confidence false detections; second, an alarm threshold is used to assess the temporal continuity of high-confidence detections. An alarm is only triggered when an EM target is continuously detected across multiple consecutive frames, effectively eliminating transient false alarms caused by lighting changes or passenger belongings.

The Execution Layer translates control commands from the Decision Layer into physical alarm actions. Once an EM is detected, the system activates voice broadcasts (e.g., Electric motorcycles are prohibited from entering elevators) and flashing LED visual alerts to realize combined audio-visual warnings. Meanwhile, key event information, including timestamps, target categories, and captured snapshots, is automatically saved to a log database. This forms a full-cycle management workflow of Perception–Decision–Execution–Logging, supporting post-event traceability and system optimization.

3.2 Core Algorithm Selection and Improvement

The YOLOv9 series has been widely adopted for various object detection tasks due to its outstanding accuracy and real-time performance. For edge deployment, model lightweighting is essential to reduce computation and power consumption. A major challenge for lightweight deep networks is severe information loss during forward propagation. YOLOv9 addresses this problem by introducing Programmable Gradient Information (PGI) and reversible functions to retain effective feature information during training and inference. The GELAN combines the advantages of Cross Stage Partial Network (CSPNet) and Efficient Layer

Aggregation Network (ELAN). It achieves a good balance among parameter quantity, computational complexity, detection accuracy, and inference speed, and is compatible with diverse edge devices [20]. To further accelerate inference, we replace standard convolutions in GELAN with DSConv. DSConv decomposes standard convolution into two independent sequential operations: depthwise convolution for spatial feature learning and pointwise convolution for channel feature fusion, improving parameter efficiency and detection performance.

3.2.1 YOLOv9 Model Structure

The overall structure of the YOLOv9 model adopted in this system is shown in Figure 2, which follows the classic three-level architecture of "Backbone-Neck-Head" commonly adopted in lightweight collaborative detection frameworks for resource-constrained edge scenarios [21] and is specifically optimized for the EM detection task in elevator scenarios:

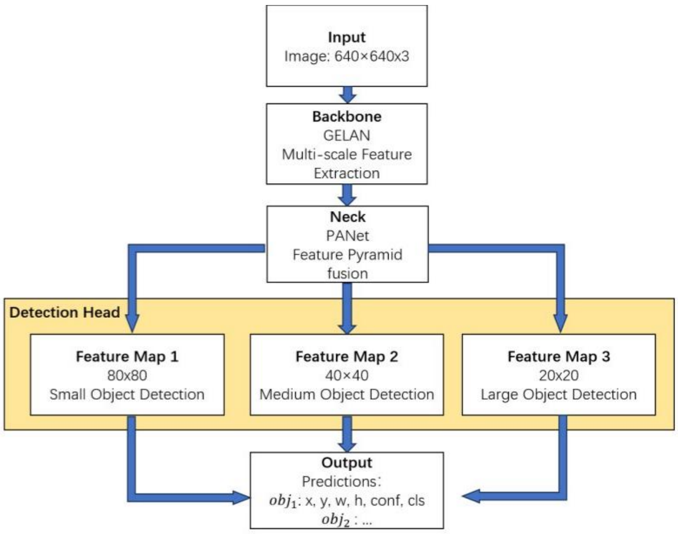


Figure 2. Schematic diagram of the YOLOv9 network architecture for EMs detection.

Backbone: Composed of stacked GELAN modules, it extracts and abstracts multi-scale feature maps layer by layer. Benefiting from GELAN's high efficiency, the backbone maintains strong feature representation under edge computing constraints.

Neck: Built upon an improved Path Aggregation Network (PANet) and Feature Pyramid Network (FPN) [23], following similar feature fusion enhancement strategies validated in traffic scenario object detection tasks [22]. Its core function is to perform top-down and bottom-up bidirectional feature fusion on multi-scale feature maps extracted by the backbone network. This fusion process combines the rich semantic information of deep features with the precise spatial location information of shallow features, generating a set of enhanced multi-scale fused feature maps suitable for detecting targets of different sizes. This effectively improves the model's detection capability for targets of varying scales inside elevators (e.g., the entire EM and local components such as wheels and batteries).

Detection Head: A decoupled head is adopted to separate classification and bounding box regression into two independent branches. This design mitigates task conflicts and further improves classification and localization accuracy, guaranteeing reliable decision-making for subsequent alarm modules.

PGI is a core component of YOLOv9 that resolves three common bottlenecks in deep network training: information compression in forward propagation, gradient attenuation or distortion in backpropagation, and error accumulation caused by traditional deep supervision. PGI introduces auxiliary branches during training to enrich gradient information, while bringing no extra computational overhead during inference. It achieves effective gradient guidance and information retention simultaneously.

The design of PGI is based on two theoretical foundations. First, the information bottleneck principle, which demonstrates that the mutual information between input data and transformed features undergoes irreversible and continuous decay through successive nonlinear transformations in the network, formalized as:

$$F(I,I) \geq F(I,f(I)) \geq F(I,g(f(I))) \tag{1}$$

where F denotes mutual information, I represents the input data, and f and g are nonlinear transformation operations such as convolution and activation. Severe loss of task-critical information leads to biased gradient signals, driving the network to learn incorrect input-target correlations, and ultimately resulting in slow convergence, poor generalization, and weak small object detection performance.

A reversible function ensures complete information retention:

$$I = R(r(I)) \tag{2}$$

where r is a reversible function and R is its inverse transformation. This transformation ensures complete information retention with invariant mutual information, fundamentally avoiding forward information attenuation and generating unbiased, reliable gradient signals.

3.2.2 Lightweight GELAN enhanced with DSConv

To enhance the model's adaptability to diverse deployment environments while preserving high detection accuracy and real-time performance, GELAN is adopted in this system. This network integrates the design philosophies of the Cross Stage Partial Network (CSPNet) and Efficient Layer Aggregation Network (ELAN), drawing on lightweight design principles similarly explored in resource-constrained detection tasks such as X-ray security inspection [25], forming a generic, flexible, and efficient layer aggregation paradigm with the following key attributes. It enables the efficient transmission and fusion of gradient and information flows across multiple network layers, and its integrated DOWN module [26]–comprising a pooling layer and a convolution layer—undertakes feature map downsampling. This operation compresses the spatial resolution to reduce the computational complexity of subsequent network layers, while expanding the number of feature channels to boost feature representation capability, thereby achieving effective capture of the target's cross-scale contextual information.

To realize model lightweighting and improve parameter utilization and computational efficiency, DSConv is introduced into the framework. DSConv consists of two sequential components: depthwise convolution and pointwise convolution. Depthwise convolution is dedicated to extracting spatial features; its convolution kernels operate in a single-channel mode, where convolution is performed on each channel of the input feature map to generate an output feature map with the same number of channels as the input. Given an input feature map with height H, width W, and M input channels (equivalently, M depthwise convolution kernels, one per channel), and performing 3×3 convolution on each channel to extract spatial features independently without channel fusion, the computational cost F_d of depthwise convolution is given by:

$$F_d = H \times W \times M \times 3 \times 3 \tag{3}$$

Since depthwise convolution only outputs feature maps with the same channel number as the input, it fails to capture cross-channel feature correlations and may lead to insufficient information representation. To address this limitation, pointwise convolution is applied subsequently. Pointwise convolution (PWConv) essentially implements channel-wise feature fusion and dimensionality expansion via 1×1 convolution kernels to extract cross-channel features. Assuming N pointwise convolution kernels are used, the computational cost F_p of pointwise convolution is calculated as:

$$F_p = H \times W \times N \times 1 \times 1 \tag{4}$$

For comparison, the computational cost of standard convolution with the same input/output configuration is given by:

$$F_{std} = H \times W \times M \times N \times 3 \times 3 \tag{5}$$

The ratio of DSConv's computational cost to that of standard convolution is thus:

$$\frac{F_d + F_p}{F_{std}} = \frac{HWM \times 9 + HWN}{HWMN \times 9} = \frac{1}{N} + \frac{1}{9} \tag{6}$$

The total computational cost of DSConv is thus only $(1/N + 1/9)$ that of standard convolution, achieving significant computational reduction.

3.2.3 System Integration and Alarm Subsystem Design

The alarm subsystem adopts a dual-evidence fusion decision framework to balance detection sensitivity and false alarm rate in complex elevator environments, drawing on multi-frame information fusion principles that combine per-frame detection confidence with temporal consistency across consecutive frames to improve detection reliability [27]. Different from conventional single-threshold strategies, it fuses two types of evidence: (1) Perception evidence: real-time detection confidence and category probability from the enhanced YOLOv9 model; (2) Temporal evidence: target presence continuity across consecutive frames. A dual-threshold logic is implemented. The system triggers intervention only when both evidence sources meet thresholds, effectively suppressing transient interference-induced false alarms.

Let $c \in [0, 1]$ denote the detection confidence of an EM target in the current frame and let $T = 0.5$ be the preset confidence threshold. The frame-level detection result c is defined as a binary variable: the target is classified as detected when its confidence meets or exceeds the threshold, and is regarded as absent otherwise. The binary decision rule is formulated as:

$$b = \begin{cases} 1, & c \geq T \\ 0, & c < T \end{cases} \tag{7}$$

The alarm system determines whether to trigger an alert based on the warning coefficient within the time window. The alarm is triggered if three consecutive frames detect an EM. Otherwise, the system calculates a warning coefficient to determine whether to initiate an alert. The warning coefficient W is calculated as:

$$\begin{aligned} W &= w_F + w_1 + \cdots + w_N \\ &= (1 - \alpha - \alpha^N)C_F + \alpha C_{F-1} + \cdots + \alpha^N C_{F-N} \end{aligned} \tag{8}$$

where C_F represents the binary detection result of the current frame, C_{F-N} corresponds to the result of the N-th preceding frame, and w_N is the warning coefficient of the N-th historical frame. The warning confidence in detecting an EM is calculated by assigning a weight to the confidence of each frame. The weights decay exponentially with time distance: the weight of the immediately preceding frame is α, and the weight of the N-th previous frame is α^N. When the warning coefficient exceeds the warning threshold ($\varepsilon = \alpha$), an alarm is activated. The weight α is empirically set to 0.4 based on experimental validation.

$$\text{Alarm} = \begin{cases} 0, & W < \varepsilon \\ 1, & W \geq \varepsilon \end{cases} \tag{9}$$

Upon alarm confirmation, a synchronized multi-modal output fusion strategy is executed to ensure reliable cross-stakeholder dissemination. This standardized protocol unifies three complementary channels with sub-500ms latency: (1) Audio: frequency-modulated warning tone for noisy environments; (2) Visual: 2 Hz flashing red GUI indicator, bold "ALARM ACTIVE" label, and thick red bounding box with confidence overlay; (3) Textual: structured event information (timestamp, category, confidence, coordinates, frame index) recorded simultaneously in the interface for audit and traceability.

An adaptive temporal fusion module with 5-second configurable cooldown further minimizes false alarms. It fuses historical alarm events and current detection results within a sliding time window to distinguish persistent threats from transient false positives. During cooldown, detection continues but repeated alarms are suppressed to avoid continuous disturbances. A dedicated "Alarm Test" function facilitates pre-deployment verification of all output channels.

4 Experiments and Results

4.1 Dataset Construction and Preprocessing

To address the scarcity of dedicated datasets for EMs detection in elevator scenarios, a large-scale, high-quality dataset for EMs detection in elevator cabins is constructed in this paper. The dataset contains approximately 2,000 images collected from multiple sources, including on-site shooting in actual elevators, web crawling, and publicly available object detection datasets. It comprehensively covers various real-world elevator conditions: different target positions inside the elevator (e.g., doorway, corner), diverse lighting conditions (bright, dim), various target states (full visibility, occlusion by passengers or luggage, human-vehicle interaction), and different brands, models, and colors of EMs. This ensures the diversity and representativeness of the dataset, laying a solid data foundation for model training and verification. All images are meticulously annotated using Labelme [28], with two annotated object categories: electric motorcycle and others (e.g. bike, people). Figure 3 illustrates representative samples of different target positions within the elevator cabin (e.g., doorway, corner), while Figure 4 presents the diverse lighting conditions captured in the dataset, ranging from bright daylight to dim nighttime illumination.

4.2 Model Training and Optimization Strategies

Model training is conducted with the following configuration and optimization strategies. The input image size is fixed at 640×640 pixels. The model is trained for a total of 150 epochs, with an early stopping strategy (patience = 10) applied to prevent overfitting. To stabilize the training process, the Stochastic Gradient Descent (SGD) optimizer is employed with a momentum of 0.937 and a weight decay of 0.0005. The initial learning rate is set to 0.01, and a cosine annealing schedule is used for dynamic learning rate adjustment, with a warmup phase to accelerate convergence. The key hyperparameter configurations and their optimization rationale are summarized in Table 1.

To enhance the model's robustness and generalization ability, a systematic data augmentation strategy is implemented during training, including mosaic augmentation, random affine transformation (scaling, translation, rotation, shearing), and color space perturbation (adjustment of hue, saturation, and brightness). These augmentation techniques effectively simulate various challenging imaging conditions encountered in real elevator environments,

Table 1. Comparison of training hyperparameter configurations.

Parameter	Baseline Configuration	Optimized Configuration	Rationale for Tuning
Epochs	100	150	Extended for better convergence on complex features
Batch size	16	10	Adjusted for GPU memory constraints (8GB)
Early stopping patience	100	10	Aggressive prevention of overfitting
Label smoothing	0.0	0.1	Enhanced model generalization
Optimizer	SGD	SGD (with comparative experiments)	Retained after optimizer ablation study
Data path	Absolute path (E:/...)	Relative path (ROOT/...)	Improved portability and reproducibility

Table 2. Performance evaluation results of the system under typical scenarios.

Typical Scenario	Number of Images	mAP@0.5	Precision	Recall	Average Processing Time (ms)
A: Complete Targets	500	0.952	0.945	0.968	28
B: Human-Vehicle Coexistence	400	0.915	0.983	0.902	35
C: Partial Occlusion	300	0.887	0.905	0.872	32

enabling the model to learn more robust target features. Furthermore, label smoothing ($\varepsilon = 0.1$) is applied to the classification loss to mitigate the model's overconfidence in training labels, following the regularization technique originally introduced for deep convolutional architectures [30]. Exponential Moving Average (EMA) is also utilized to update the model weights, a strategy adopted in prior real-time detection frameworks such as YOLOv4 to facilitate convergence to a more stable and generalizable optimal solution [29], with optimizer configuration further informed by decoupled weight decay regularization practices [31].

4.3 Comparative Analysis

To comprehensively evaluate the performance of the proposed system, we conducted experiments on a computer equipped with an RTX 3060 mobile graphics card. The constructed dataset of approximately 2,000 images is divided into training, validation, and test sets at a ratio of 8:1:1. The model is trained with an input image size of 640×640 for 150 epochs, using the AdamW optimizer and data augmentation techniques including mosaic and random affine transformation. Label smoothing and EMA are applied throughout the training process. System performance is evaluated on three specially designed test subsets, which cover the main typical scenarios of EM entry into elevators.

Adopting mean average precision (mAP@0.5), precision, and recall to evaluate the algorithm performance, and measuring the average processing time per frame. The detailed performance evaluation results of the system on the three test subsets are shown in Table 2. Figure 5 shows the detection results for typical scenarios.

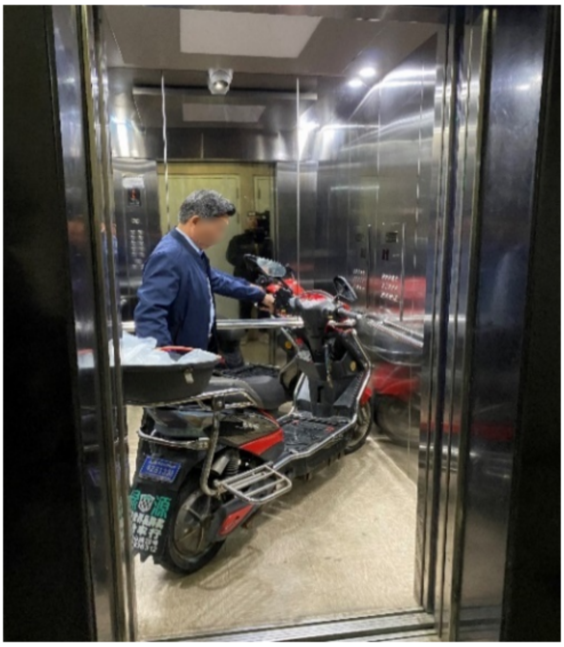
In Scenario A (where the target is fully visible), the EM target is fully visible, unobstructed, and free from human interference, representing ideal detection conditions for evaluating the upper limit of the model's basic detection capabilities. As shown in Figure 5(A), the target bounding box closely matches the actual contour of the target, with no missed detections or false positives, and no false responses in the foreground area. This achieves precise classification and localization of the target. In this subset, the mAP@0.5 reached 0.952, with a single-frame processing time of only 28 ms. It is the scenario with the highest accuracy and fastest inference speed among the three, validating the model's fundamental detection performance and real-time processing capabilities under ideal conditions.

In Scenario B (Human-Vehicle Coexistence), EM and pedestrians overlap spatially and partially obstruct each other, representing one of the most common real-world scenarios encountered in elevators. In this subset, the mAP@0.5 reaches 0.915, with a precision rate as high as 0.983 and a processing time of 35

Table 3. Electric vehicle detection and alarm results.

Frame	Previous 4 Frames	Previous 3 Frames	Previous 2 Frames	Previous 1 Frames	Current Frame
Confidence Level	0	0.84	0.43	0.83	0
Electric Vehicle	0	1	0	1	0
Warning Coefficient	0	0.064	0	0.4	0

a: door



b: corner

Figure 3. Different positions within the elevator.

a: dim

b: bright

Figure 4. Diverse lighting conditions.

ms per frame. Even in complex human-vehicle interaction scenarios, there is only minimal accuracy degradation, while maintaining a high precision rate with near-zero false positives. This is attributed to PGI's multi-level auxiliary information module, which generates supervision signals by fusing full-scale target information to guide phased training of the main and branch networks. This approach avoids feature bias caused by single-level supervision, enabling the model to learn the core features of electric vehicles while accurately distinguishing between pedestrians and targets in human-vehicle coexistence scenarios, ensuring that targets are not misclassified as background due to pedestrian occlusion.

In Scenario C (Partial Occlusion), the EM target is severely obscured by people inside the elevator and the scooter's windshield, with less than 50% of the target visible and feature information severely incomplete. This is the most challenging high-frequency scenario in the elevator. However, in this scenario, the accuracy reaches 0.887, with a processing time of 32 ms per frame. This is because, during training, PGI guides the backbone network to learn the core features of the e-bike target rather than its complete contour features,

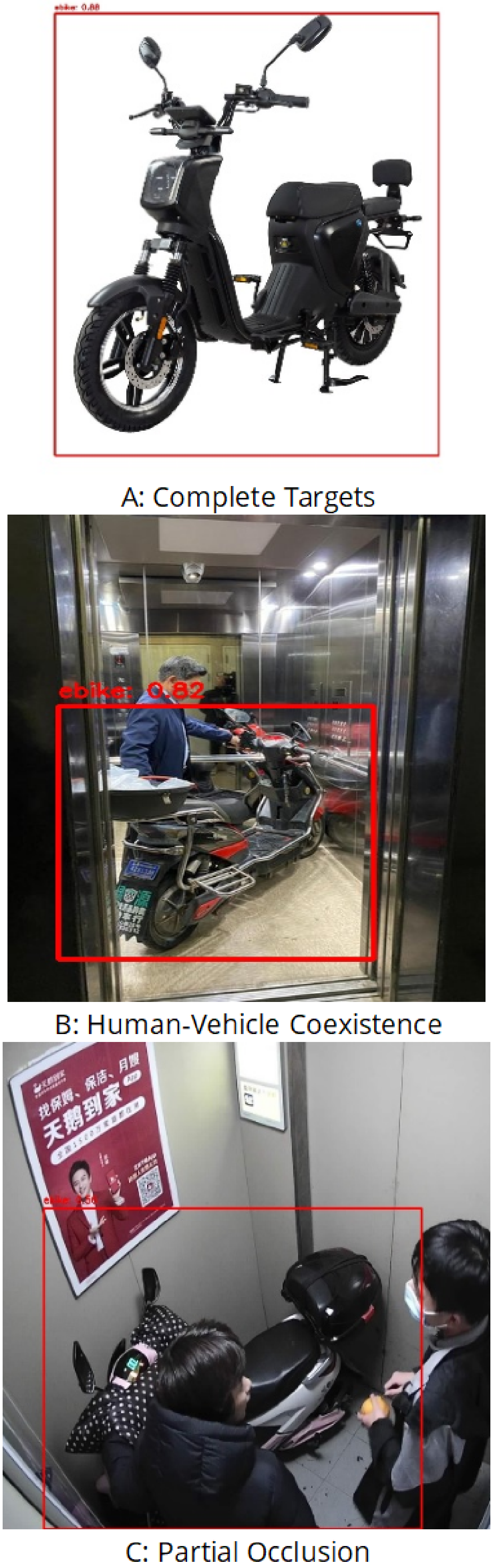


Figure 5. The results of the system under typical scenarios.

enabling the model to identify the EM target from incomplete local features.

To quantitatively evaluate the alarm performance of the proposed multi-source information fusion alarm system, we conducted validation tests in a real elevator cabin, with the size of the temporal sliding window set to 5 frames. Due to frequent occlusion by passengers, EM targets cannot be consistently detected across all consecutive frames, resulting in intermittent, discontinuous detection outputs.

The conventional single-frame threshold strategy triggers alarms solely based on the detection confidence of the current frame. Such a mechanism would produce frequent on-off intermittent alarms in practical deployment, severely degrading system usability. In contrast, the proposed dual-threshold fusion mechanism incorporates temporal continuity information by assigning weighted values to detection results of historical frames. As presented in Table 3, even when no EM is detected in the current frame, the system still outputs one stable valid alarm by leveraging detection evidence from the 1st and 3rd preceding frames. The result confirms that the proposed method effectively suppresses redundant intermittent alarms caused by transient occlusion, reduces false alarm rates in complex elevator scenarios, and improves the overall reliability and operational efficiency of the alarm system.

In summary, the enhanced YOLOv9 model achieves a balance between high accuracy and high real-time performance in ideal scenarios with fully visible targets. It maintains exceptional detection robustness in high-frequency, challenging elevator scenarios—such as those involving both people and vehicles or partial occlusions— low false positives, meeting the engineering requirements for detecting and alerting against the prohibited entry of electric vehicles into elevators.

5 Conclusion

This study focuses on the critical fire safety risks posed by the illegal entry of EMs into building elevators, and systematically investigates the core challenges inherent to confined elevator cabin environments: frequent target occlusion and highly variable lighting conditions. A dedicated custom dataset comprising approximately 2,000 images is compiled for the detection task. It comprehensively covers diverse target positions within elevator cabins, bright and dim lighting conditions, human-vehicle interaction scenarios, and typical occlusion caused by passengers or carried belongings, providing a robust data foundation for model robustness training and comprehensive performance evaluation. At the algorithmic level, we integrate the built-in PGI training mechanism of YOLOv9 with a DSConv-enhanced GELAN backbone. The proposed model achieves a well-balanced trade-off between detection accuracy and real-time performance in complex elevator scenarios, yielding mAP@0.5 scores of 0.952, 0.915, and 0.887 across three representative test scenarios. The per-frame inference latency remains consistently below 35 ms, fully meeting the performance requirements for real-time detection on edge computing devices. Building on the optimized detection model, a complete closed-loop EM detection and alarm system is developed and validated. The system adopts a hierarchical modular design with a three-layer architecture (Perception Layer, Decision Layer, and Execution Layer), and implements a multi-source information fusion framework for real-time EM detection and alerting. It incorporates a dual-threshold false alarm suppression mechanism based on confidence and temporal fusion, as well as multimodal audio-visual-log alarm outputs. This design establishes a stable full-cycle closed loop spanning risk perception, fusion-aided decision-making, alarm execution, and event traceability, effectively balancing the system's detection sensitivity and false alarm rate. The system delivers reliable alarm performance and exhibits strong engineering applicability for practical elevator monitoring deployments.

Nevertheless, the scenarios evaluated in this study are largely limited to relatively well-controlled lighting conditions and typical occlusion scenarios in elevator cabins. To further improve the generalization capability and environmental robustness of the system, future research will prioritize systematic validation experiments under more extreme and challenging conditions, including severe target occlusion, ultra-low illumination, and drastic lighting fluctuations. This will enable the system to maintain stable and reliable detection performance in highly complex, unconstrained real-world deployment scenarios.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

DeepSeek-R1 was used for proofreading the entire manuscript and for checking the formatting of references. DeepL Translator was used for translating references originally published in Chinese into English. No generative AI was used for content generation, experimental design, data analysis, or interpretation of results. The authors take full responsibility for the content, accuracy, and originality of this manuscript.

Ethical Approval and Consent to Participate

The elevator surveillance footage used in this study was obtained with formal authorization from the property management companies of the participating residential buildings, in accordance with their existing security camera operation policies, which include standard notice to residents that footage may be used for safety research and system improvement purposes. The research team did not independently install any recording equipment; all data were extracted from pre-existing, authorized building surveillance systems. No additional personal information beyond visual imagery was collected, and the study does not involve any direct interaction with or recruitment of human participants. Individuals incidentally captured in the published figures have been informed through standard building notices that surveillance recordings may be used for research purposes; where feasible, facial regions have been obscured in the published figures to further protect individual privacy.

References

[1] Sheng-ping, T., Xu-hong, H., Nan, Z., Wei-hao, X., Yan-yi, C., & Cun-xin, Z. H. U. (2022, November). Research on two-wheeled vehicle entry ban system based on deep learning. In *International Conference on Advanced Intelligent Systems and Informatics* (pp. 224-233). Cham: Springer International Publishing. [CrossRef]

[2] Li, L., Liu, B., Zheng, W., Wu, X., Song, L., & Dong, W. (2022). Investigation and numerical reconstruction of a full-scale electric bicycle fire experiment in high-rise residential building. *Case Studies in Thermal Engineering*, *37*, 102304. [CrossRef]

[3] Wang, Y., Zhang, Y., & Yu, Z. (2022, August). Fire risk assessment of electric bicycles charging facilities in old urban communities. In *2022 Power System and Green Energy Conference (PSGEC)* (pp. 447-451). IEEE. [CrossRef]

[4] Han, Y., & Wang, W. (2025). Identification System for Electric Bicycle in Compartment Elevators. *Electronics, 14*(13), 2638. [CrossRef]

[5] Li, T., Lei, L., Wang, Z., Shi, P., & Wu, Z. (2025). An efficient improved YOLOv10 algorithm for detecting electric bikes in elevators. *Electronic Research Archive, 33*(6), 3673-3698. [CrossRef]

[6] Huang, J., Lin, G., Dong, Z., & Sun, F. (2025, June). Research on Real-time Object Detection Algorithms for Elevator Intelligent Monitoring Scenarios. In *Proceedings of the 2025 2nd International Conference on Computer and Multimedia Technology* (pp. 229-235). [CrossRef]

[7] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016, June). You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 779-788). IEEE Computer Society. [CrossRef]

[8] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023, June). YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7464-7475). IEEE. [CrossRef]

[9] Zhang, X., Yuan, D., Sun, H., Yu, L., Guo, W., & Feng, X. (2025, June). YOLOv5-Based Detection of Electric Vehicles and Batteries in Residential Elevators for Fire Hazard Prevention. In *2025 5th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA)* (pp. 1-8). IEEE. [CrossRef]

[10] He, Y. (2024). Automatic detection of electric motorcycle based on improved YOLOv5s network. *Journal of Electrical and Computer Engineering, 2024*(1), 4889707. [CrossRef]

[11] Saleh, K., Szénási, S., & Vámossy, Z. (2021, January). Occlusion handling in generic object detection: A review. In *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)* (pp. 000477-000484). IEEE. [CrossRef]

[12] Zhu, Z., Cao, J., Hao, T., Zhai, W., Sun, B., Jia, G., & Li, M. (2020). Highly secure edge-intelligent electric motorcycle management system for elevators. *Journal of Cloud Computing, 9*(1), 41. [CrossRef]

[13] Liu, P., Li, Y., Yuan, L., & Wang, Y. (2026). A Lightweight Method for Detecting the Behavior of Electric Bicycle Entering Elevators Based on Deformation Multi-scale Collaboration. In *International Conference on Computing and Pattern Recognition* (pp. 393-403). Springer, Singapore. [CrossRef]

[14] Lin, Y., Chen, X., Zhong, W., & Pan, Z. (2021, March). Online detection system for electric bike in elevator or corridors based on multi-scale fusion. In *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (pp. 31-34). IEEE. [CrossRef]

[15] Wan, Y., Hu, Y., Li, X., Song, Z., & Hou, T. (2024, November). Improved Algorithm of Elevator Blocking System Based on YOLOv8. In *2024 12th International Conference on Information Systems and Computing Technology (ISCTech)* (pp. 1-5). IEEE. [CrossRef]

[16] Zhao, Z., Li, S., Wu, C., & Wei, X. (2023). Research on the rapid recognition method of electric bicycles in elevators based on machine vision. *Sustainability, 15*(18), 13550. [CrossRef]

[17] Zhu, K., Dai, J., & Gu, Z. (2024). Dynamic tracking method based on improved deepsort for electric vehicle. *World Electric Vehicle Journal, 15*(8), 374. [CrossRef]

[18] Wang, C. Y., Yeh, I. H., & Mark Liao, H. Y. (2024, September). Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision* (pp. 1-21). Cham: Springer Nature Switzerland. [CrossRef]

[19] Ikmel, G., & El Amrani, E. L. (2024, June). Performance analysis of YOLOv5, YOLOv7, YOLOv8, and YOLOv9 on road environment object detection: comparative study. In *2024 international conference on ubiquitous networking (UNet)* (Vol. 10, pp. 1-5). IEEE. [CrossRef]

[20] Wang, L., Letchmunan, S., & Xiao, R. (2024). Gelan-SE: squeeze and stimulus attention based target detection network for gelan architecture. *IEEE Access, 12*, 182259-182273. [CrossRef]

[21] Chen, Z., Fan, K., Ye, J., Xu, Z., & Wei, Y. (2025). A Lightweight Multi-Module Collaborative Optimization Framework for Detecting Small Unmanned Aerial Vehicles in Anti-Unmanned Aerial Vehicle Systems. *Drones, 10*(1), 20. [CrossRef]

[22] Li, Y., Zheng, Q., Xu, S., Wang, P., Wang, Y., Song, Z., & Guo, M. (2025). YOLO-EDGE: an object detection algorithm for traffic scenarios. *The Journal of Supercomputing, 81*(7), 840. [CrossRef]

[23] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018, June). Path Aggregation Network for Instance Segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8759-8768). IEEE. [CrossRef]

[24] KIM, T., Sun-Been, P., Yu-Jeong, J., & Da-Eun, L. (2024). A study on the elevator system using real-time object detection technology YOLOv5. *Journal of Internet of Things and Convergence, 2024, 10*(2), 103–108. [CrossRef]

[25] Yu, X., Chen, C., Cheng, S., & Li, J. (2025). A lightweight detection method for X-ray security inspection based on YOLOV8. *Applied Soft Computing*, 113822. [CrossRef]

[26] Su, J., Yang, M., & Tang, X. (2024). Integration of ShuffleNet V2 and YOLOv5s networks for a lightweight object detection model of electric bikes

within elevators. *Electronics, 13*(2), 394. [CrossRef]

[27] Long, X., Hu, S., Hu, Y., Gu, Q., & Ishii, I. (2019). An FPGA-based ultra-high-speed object detection algorithm with multi-frame information fusion. *Sensors, 19*(17), 3707. [CrossRef]

[28] Torralba, A., Russell, B. C., & Yuen, J. (2010). Labelme: Online image annotation and applications. *Proceedings of the IEEE, 98*(8), 1467-1484. [CrossRef]

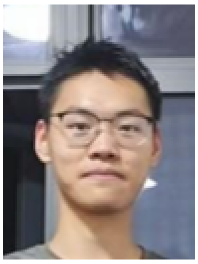
[29] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*. [CrossRef]

[30] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016, June). Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2818-2826). IEEE. [CrossRef]

[31] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. [CrossRef]

Qin Luo is currently an undergraduate student majoring in Electronic Information Engineering at the College of Electronic Science and Technology, National University of Defense Technology, Changsha, China. His research interests focus on computer vision, deep learning-based object detection, and intelligent safety systems. (Email: 1903516237@qq.com)

Yibo Sun is currently an undergraduate student majoring in Electronic Information Engineering at the College of Electronic Science and Technology, National University of Defense Technology, Changsha, China. His research interests include real-time video analysis, deep learning model optimization, and embedded system development. (Email: 3085495398@qq.com)



Pengxiang Lu is currently an undergraduate student majoring in Electronic Information Engineering at the College of Electronic Science and Technology, National University of Defense Technology, Changsha, China. His research interests lie in edge computing, intelligent alarming systems, and human-computer interaction design. (Email: 1649887769@qq.com)

Zhichao Yin received the MS degree in Geomatics Science and Technology from Chongqing Jiaotong University, China in 2024. Currently serves as an engineer in Key Laboratory of Automatic Target Recognition, National University of Defense Technology, P. R. China. His research interests indoor positioning, target detection and recognition, intelligent image real-time information processing, and Pedestrian Activity Recognition. (Email: 15200430340@163.com)

Moufa Hu received the BS, MS and PhD degree from National University of Defense Technology, Changsha, P. R. China, in 2001, 2003 and 2008. He is now an associate researcher in Key Laboratory of Automatic Target Recognition, National University of Defense Technology, P. R. China. He has authored or coauthored more than 40 journal and conference proceedings papers. He has been the principal investigator of 20 research papers. His current research interests include optical imaging target detection and recognition, intelligent image real-time information processing, and algorithm performance evaluation. (Email: mfhu20080630@nudt.edu.cn)