IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 April 2025, accepted 7 May 2025, date of publication 20 May 2025, date of current version 28 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3569101

APPLIED RESEARCH

License Plate Recognition for Smart Construction Sites Based on GMH-YOLO

MING LI[1,2], ZE-QUAN WANG[1,2], YU-HANG ZHAO[1,2], AND QIANG LI[3]
[1]Business School, Hohai University, Nanjing 210000, China
[2]Project Management Informatization Institute, Hohai University, Nanjing 210000, China
[3]Zhongbo Information Technology Research Institute Company Ltd., Nanjing 210001, China

Corresponding author: Ze-Quan Wang (2208080125@hhu.edu.cn)

This work was supported by the Fundamental Research Funds for the Central Universities under Grant B230207087.

ABSTRACT With the rapid urbanization, vehicle management at construction sites has become crucial for safety and logistics efficiency. However, License Plate Detection in such environments faces unique challenges, such as simultaneous detection of dual plates and body plates, strong light reflections, and muddy interferences. This paper constructs a license plate dataset, CSLPD, specifically for construction sites, containing 1495 images and 2301 license plate instances, with double and body license plates accounting for 27.2% and 25.4%, respectively. To enhance detection performance in complex environments, this paper proposes the GMH-YOLO model, which integrates an innovative Gated Multi-Head Attention mechanism into the C3k2 module of YOLO11. This lightweight gating unit adaptively allocates feature channel resources, effectively enhancing key information while suppressing background interference, making it particularly suitable for detecting multiple license plate types and partially occluded plates in complex construction site environments. Experimental results show that GMH-YOLO achieves 93.3% mAP@50 on the CSLPD dataset, outperforming YOLO11 by 1.4%. For the challenging body license plate task, detection accuracy improves from 81.3% to 87.2%, a 5.9% increase. The model maintains high real-time performance due to the optimized gating mechanism. Comparative experiments with six attention mechanism integration schemes confirm that the gated mechanism provides the best balance between feature extraction and computational efficiency, offering a high-precision, efficient solution for intelligent license plate recognition at construction sites.

INDEX TERMS License plate recognition, smart construction site, GMH-YOLO, CSLPD.

I. INTRODUCTION

With the rapid advancement of global urbanization, the demand for intelligent management of construction sites has become increasingly prominent. Construction site vehicle management [1] plays a key role in ensuring construction safety and improving logistics efficiency. As a critical component, license plate recognition technology must address unique challenges in construction site environments, despite being widely deployed in Intelligent Transportation Systems (ITS) [2] and access control management.

Recent object detection algorithms based on deep Convolutional Neural Networks (CNNs) [3] have made significant progress in solving complex visual problems. Traditional Automatic License Plate Recognition (ALPR) systems consist of three main stages: License Plate Detection (LPD) [4], Character Segmentation (CS) [5], and Optical Character Recognition (OCR) [6]. License plate detection forms the foundation of ALPR systems, directly affecting subsequent module performance.

Construction site environments often face more complex and unique conditions compared to urban traffic scenarios. First, most construction sites feature harsh conditions, with heavy dust particles continuously suspended in the air, significantly reducing visibility and accumulating on license plate surfaces, causing partial or complete character obscuration. Second, construction sites experience extreme lighting variations, causing severe glare on license plates, resulting

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhua Guo.

in dramatic contrast differences even within a single image. Furthermore, construction vehicles typically employ multiple license plate modes simultaneously: standard regulatory plates, auxiliary plates indicating vehicle type or function, and large identifiers painted on vehicle bodies to maintain visibility from multiple angles. The practicality of these identification enhancements also increases the complexity of detection, especially for body license plates with varying dimensions, positions, and visual characteristics. Finally, the continuous vibration and movement of both vehicles and camera systems in active construction environments introduce motion blur that further complicates accurate license plate recognition.

In construction environments, real-time performance and detection accuracy are equally important. On one hand, construction site vehicle management requires systems capable of processing information in real-time to ensure safety and efficiency; on the other hand, speed without accuracy fails to meet practical application requirements. Therefore, this research aims to develop a solution that effectively addresses these complex challenges while maintaining computational efficiency.

Single-stage detection algorithms represented by YOLO [7] have demonstrated high accuracy and real-time processing capabilities in object detection. While versions such as YOLOv5 [8], YOLOv7 [9], and YOLOv8 [10] have achieved remarkable results in license plate detection, existing methods still show limitations in handling dual license plate detection and vehicle license plate recognition within complex construction environments.

To address these challenges, this study introduces the Construction Site License Plate Dataset (CSLPD), specifically designed for complex construction site scenes. The dataset encompasses vehicle images with dual license plates and body license plates, including instances of clear, blurred, occluded, and defaced license plates, providing substantial real-world data support for recognition tasks.

This paper also proposes the GMH-YOLO algorithm based on YOLO11 [11], integrating an improved Gated Multi-Head self-attention (GMHSA) mechanism. This solution effectively addresses the key challenges of construction site license plate recognition: For multiple co-existing license plate types, the mechanism adaptively focuses on targets of different scales and features through the multi-head; For environmental interference issues, the gating mechanism calculates feature importance weights, enhancing key feature representations while suppressing background noise and interference information, improving model robustness under complex conditions such as strong light reflection and dust interference; To balance real-time performance and accuracy, the gated attention mechanism adopts a lightweight design, enhancing model representation capability while adding minimal computational overhead, ensuring the system meets real-time requirements for practical applications.

The main contributions of this paper include:

First, our proposed gated multi-head attention mechanism (GMH) optimizes the feature extraction process. GMH adaptively calculates attention weights through a lightweight gating unit, enabling the model to intelligently allocate resources across different channels, differentially processing key feature regions and background information. This mechanism is particularly suitable for handling standard license plates and body license plates simultaneously present at construction sites, as well as license plates partially obscured by strong light and dust.

Second, we conduct extensive experiments demonstrating GMH-YOLO's superior performance in construction site license plate detection, achieving 93.3% mAP@50 (1.4% higher than YOLO11), with vehicle body license plate detection accuracy improving by 5.9% while maintaining computational efficiency.

Finally, we provide comprehensive evaluation of six improvement methods based on the C3k module and attention mechanisms, providing new insights into attention mechanism applications in object detection.

The subsequent sections present related work and datasets (Section **II**), detail the proposed model (Section **III**), discuss experimental results (Section **IV**), and conclude with findings and future directions (Section **V**).

II. RELATED WORK

A. OBJECT DETECTION IN LICENSE PLATE RECOGNITION

In recent years, CNNs have been at the forefront of object detection. Compared to traditional manual feature extraction methods [12], [13], [14], deep learning offers abundant training data, higher efficiency, and greater flexibility. Detection algorithms can generally be divided into two categories: single-stage detection algorithms, such as SSD [15] and YOLO, and two-stage detection algorithms, such as R-CNN [16] and Faster R-CNN [17]. To ensure real-time detection, the license plate recognition field often opts for faster and lighter single-stage detection algorithms.

Luo and Liu [8] proposed a vehicle license plate recognition method based on an improved YOLOv5m and LPRNet model. The K-means++ algorithm is used to improve the matching between anchor frames and detection targets, and the DIOU loss function is employed to enhance the NMS method. Additionally, they removed the 20×20 feature map and designed a license plate recognition system based on YOLOv5m-LPRNet, achieving both real-time performance and stability.

In the study [18], the YOLO-World model was proposed to address license plate detection across diverse scenarios in various countries and regions. The model was tested on the American datasets, including Stanford Cars Dataset, Used Car Dataset, and Automobile Dataset, the Indian license plate dataset Real-time Dataset, and the large Chinese dataset CCPD. It achieved notable accuracy and real-time performance across these datasets. Chung et al. [9] proposed the YOLO-SLD model [17], an improved version of YOLOv7,

which integrated the lightweight attention module SimAM to develop SIMAM-ELAN and SIMAM-ELAN-H modules. These modules considered the correlation between spatial and channel factors, generating realistic three-dimensional weights to improve the model's convergence performance. When tested on the large public CCPD dataset, the model achieved an increase in mAP@50 by 0.47%. In the study [10], the YOLOv8 model was utilized, and a specialized GUI widget was developed specifically for YOLOv8. This widget aimed to support developers in efficiently completing YOLOv8 training and inference tasks, while also enhancing development efficiency.

Moussaoui et al. [19] proposed a high-precision license plate detection and recognition method by integrating YOLOv8 and OCR techniques. Their system first employs YOLOv8 to detect license plate regions in images, then applies the k-means clustering algorithm, thresholding techniques, and opening morphological operations to enhance the image, making characters in the license plate clearer before using OCR. The method achieved convincing results in both detection and recognition performance, reaching accuracy rates of 99% and 98%, respectively.

Recently, Ismail et al. [20] presented a dual-stage approach for real-time license plate detection and recognition on mobile security robots. Their system combines the YOLOv7x model for license plate detection with a vision transformer-based recognizer (ViTLPR). ViTLPR utilizes the self-attention mechanism to read character sequences on license plates, without requiring character segmentation, by directly treating license plate recognition as a sequence labeling problem. Extensive experiments on their self-built PGTLP-v2 dataset and five other benchmark datasets demonstrated that their proposed ALPR system outperforms existing baseline methods while maintaining efficient inference speed.

B. YOLO MODEL NETWORK

The YOLO framework has evolved significantly through continuous innovations from the Ultralytics team and research community. YOLOv5 integrates CSPNet [21] and Path Aggregation Network (PAN) [22] concepts, employing the CSP structure to optimize computational efficiency while maintaining accuracy. YOLOv8 marks another significant advancement with its C2f module replacing the traditional C3 module, coupled with a RepVGG-inspired [23] decoupled head design and refined loss functions incorporating Variable Focal Loss (VFL) and Distributed Focal Loss (DFL).

YOLO11 [11], the latest iteration from Ultralytics, introduces several breakthrough improvements over YOLOv8. The model incorporates C3k2 and C2PSA modules while adopting YOLOv10's [24] detection head principles. Through the integration of Deep Separable Convolution (DWConv), YOLO11 achieves enhanced performance in complex scenes and small target detection while maintaining a lightweight architecture. This balanced design makes it versatile across platforms, from edge devices to high-performance systems. Each component's contribu-tion will be analyzed in the experimental section.

As shown in Figure 1, YOLO11 is divided into three parts: Backbone Network (Backbone), Neck Network (Neck) and Head Network (Head). In the Backbone Network, YOLO11 introduces the C3k2 module, as shown in Figure 2. This module is an improvement based on C2f, and the convolution content of the C3 module can be controlled by setting True or False. In addition, the Cross-stage Local Network with Parallel Spatial Attention(C2PSA) module is added to the end of the backbone network to enhance the spatial attention mechanism.

In addition, the Neck still adopts the Path Aggregation Network (PANet) structure, which enhances the context representation ability of the model through multi-scale feature fusion. It also improves the bottom-up feature propagation path to provide more comprehensive and high-quality feature information for the head network.

In the Head, YOLO11 uses DWConv to significantly reduce the computational cost and number of parameters while speeding up feature processing. The introduction of DWConv not only improves the inference efficiency of the model, but also further optimizes the accuracy of multi-scale prediction. By outputting multi-level feature maps, the head network enables YOLO11 to achieve new performance levels in target detection in complex scenarios while maintaining a lightweight design.

C. ATTENTION MECHANISM

1) ATTENTION MECHANISMS IN RELATED VISUAL TASKS

The challenges in license plate recognition in construction site environments, such as distinguishing targets from complex backgrounds, handling partial occlusions, and adapting to varying lighting conditions, are significantly similar to the challenges in saliency detection and interactive image segmentation tasks. The innovative methods in these related fields provide valuable insights into addressing license plate recognition challenges in complex environments.

Chen et al. in [25] explored the issue of ''feature spatial independence,'' pointing out that this is one of the root causes of poor performance of saliency detectors in complex areas (such as object boundaries). This observation inspires us to consider: in the complex environment of construction sites, is the blurred boundary problem in license plate detection also derived from similar feature independence? This perspective is valuable for addressing plates with similar textures, partial occlusions, or uneven lighting.

Another related work is ''Gaussian Dynamic Convolution for Efficient Single-Image Segmentation'' [26] proposed by Sun et al., which is inspired by the human visual system and simulates dynamic receptive field mechanisms. They point out that traditional convolution operations are limited by fixed receptive fields and struggle to adapt to feature extraction of

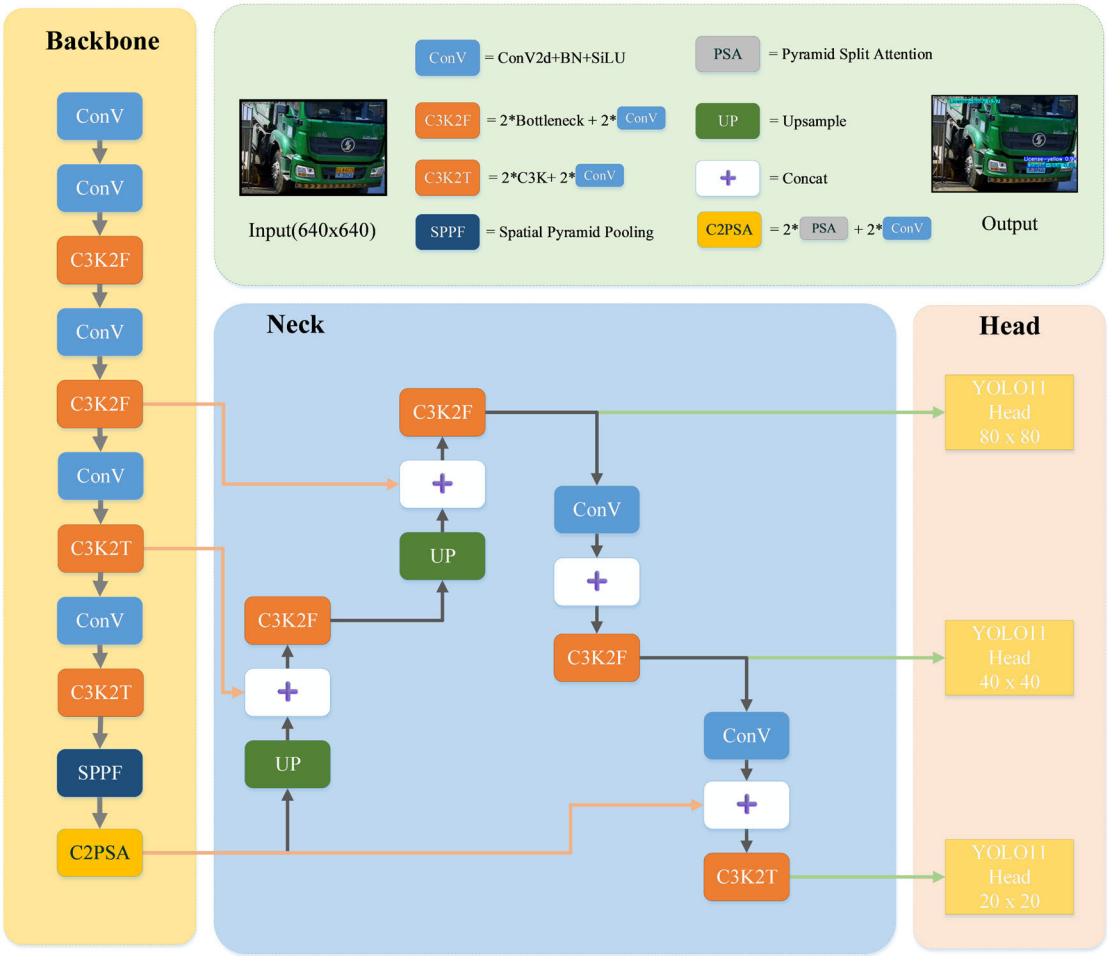

FIGURE 1. Original YOLO11 overall framework.

targets at different scales. In construction site scenes, the scale difference between standard license plates and body license plates is significant, which is highly similar to the problem solved by GDC. GDC can simultaneously capture features at multiple scales by randomly selecting spatial sampling areas from a Gaussian distribution to form dynamic convolution kernels. This multi-scale feature capture capability is particularly valuable for scenarios that process different types of license plates (standard plates, double plates, body plates) simultaneously.

Although the GDC offers significant advantages in processing multi-scale features, our research opted for an attention mechanism-based approach rather than directly adopting GDC. This is primarily because license plate detection in practical applications requires balancing real-time performance and accuracy, while GDC's random sampling process would increase computational burden. Nevertheless, the core concept of GDC—breaking the limitation of fixed receptive fields in traditional convolution and establishing more flexible feature extraction mechanisms, and it inspired our subsequent design of the gated multi-head attention module. Our GMH-YOLO draws on the concept of dynamic feature aggregation, adaptively adjusting weights of different features through a gating mechanism, enabling effective detection of license plates at different scales while maintaining computational efficiency.

2) ATTENTION MECHANISM IN OBJECT DETECTION

In object detection, attention mechanisms have yielded superior detection performance. For instance, In YOLO-SLD [9] for license plate recognition, incorporating the lightweight SimAM [27] module improved detection results compared to the original version while maintaining a lightweight and efficient model. In remote sensing image detection, Zhang et al. [28] introduced a custom FFCA attention module to enhance the extraction of shallow features and spatial information, optimizing the model's ability to extract features from targets of different scales and significantly improving the prediction of small targets. In the application of monitoring fish diseases, Cai et al. [29] integrated the NAM module into YOLOv7's ELAN. By using channel and spatial attention modules, they suppressed less significant features in the dataset, achieving

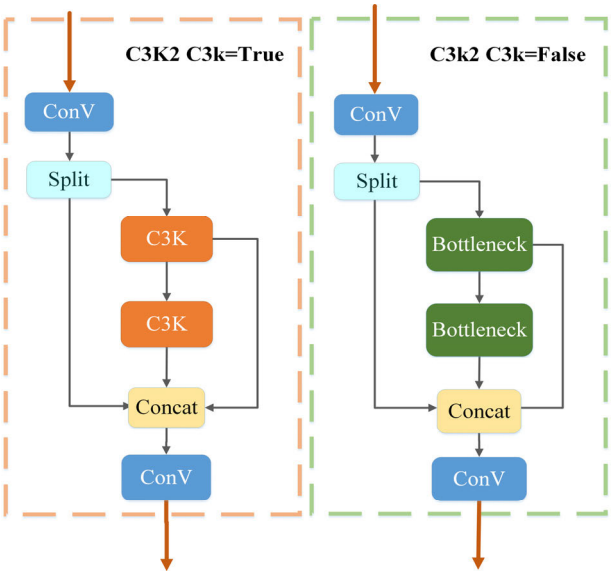

FIGURE 2. C3k2 module diagram. (The left side represents C3k = True, and the right side represents C3k = False).

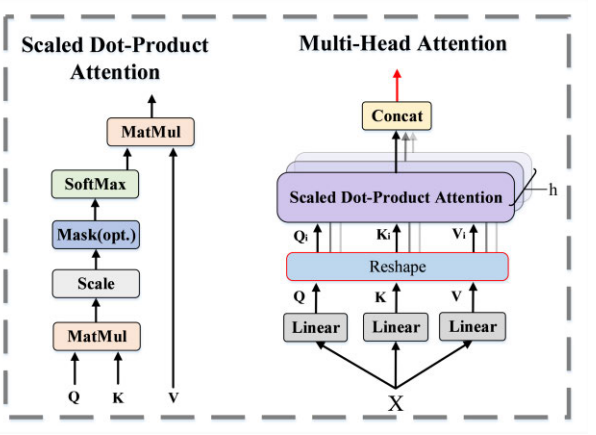

FIGURE 3. Principle diagram of MHSA attention mechanism (the red part indicates improvement over the original MHSA).

more accurate and efficient detection methods. In papers [10], [30], [31], the introduction of MHSA demonstrated high accuracy, better robustness for detecting different targets, and real-time performance.

For detecting multi-license plate identifiers (e.g., body identifiers and standard plates) on construction site vehicles, this study adopts MHSA as a feature enhancement module. Compared to traditional attention mechanisms such as CBAM and SimAM, MHSA offers the following significant advantages.

Multi-Scale Target Processing. MHSA's multi-head parallel mechanism allows simultaneous processing of targets at different scales. In this task, vehicle body identifiers (focusing on character shapes, spacing, and boundary features while ignoring background color information) usually exhibit greater scale variations than standard plates. Multiple attention heads can independently focus on features of different scales, better adapting to such scale differences. In contrast, CBAM's serial channel and spatial attention structure processes features sequentially, making it less effective at building associations between multi-scale targets.

Long-Range Dependency Modeling. MHSA's multi-head mechanism directly establishes long-range dependencies. While SimAM excels in computational efficiency and model simplicity, its variance-based adaptive feature enhancement mechanism effectively highlights important features. However, for tasks like detecting body identifiers with subtle edge and character spacing features, SimAM's simplified attention calculation limits the model's ability to model complex relationships between multi-scale targets. In comparison, MHSA's multi-head mechanism simultaneously attends to multi-scale feature expressions, making it more suitable for such complex scenarios.

3) MHSA

Based on the foundational theory of Transformers, Vaswani et al. first proposed the Multi-Head Self-Attention mechanism (MHSA) [32], an innovative mechanism capable of performing attention calculations in different feature spaces simultaneously. Unlike traditional attention mechanisms, MHSA employs multiple parallel attention heads for feature extraction, with each head focusing on different representation subspaces, thereby achieving more comprehensive feature representations. This mechanism demonstrates significant advantages in vision tasks, especially in scenarios requiring simultaneous handling of multiple spatial regions and features at different scales.

Compared to the original MHSA, the improved version (in Figure 3) removes the final linear fusion layer and replaces the intermediate convolution operations with lightweight Reshape operations. This effectively reduces the model's parameter count and inference latency while improving compatibility with the characteristics and performance requirements of license plate recognition.

The working principle of the improved MHSA module can be described as follows: the input feature map $X \in R^{C\times H\times W}$ is processed through three different linear convolutions to obtain $\boldsymbol{Q}$(Query), $\boldsymbol{K}$(Key)and$\boldsymbol{V}$(Value), expressed as $\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W_q}, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W_k}, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W_v}$, where $\boldsymbol{W_q}, \boldsymbol{W_k}, \boldsymbol{W_v}$ are learnable weight matrices. Given the number of feature subspaces h, the feature map is reshaped into (h, C//h, H × W), generating features $\boldsymbol{Q_i}$, $\boldsymbol{K_i}$ and $\boldsymbol{V_i}$ for each head. Each head is responsible for a different feature subspace and is represented as:

$$\mathbf{head_i} = \text{Attention}(\boldsymbol{Q_i}, \boldsymbol{K_i}, \boldsymbol{V_i}) \tag{1}$$

The Scaled Dot-Product Attention module is then applied, which can be expressed as:

$$\text{Attention}(\boldsymbol{W_q}, \boldsymbol{W_k}, \boldsymbol{W_v}) = \text{Softmax}(\frac{\boldsymbol{W_q}\boldsymbol{W_k^T}}{\sqrt{d_k}})\boldsymbol{V} \tag{2}$$

In the formula above, the term d_k represents the dimensionality of the Query input, and $W_q W_k^T$ calculates the similarity between the Query and Key. The term $1/\sqrt{d_k}$ prevents the Softmax function from saturating (approaching 0 or very large values), ensuring smoother attention weight distribution.

Finally, all heads are concatenated along the channel dimension. The summary formula for MHSA is:

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\mathbf{head_1},\cdots,\mathbf{head_n}) \quad (3)$$

III. PROPOSED METHOD

A. ANALYSIS OF C3K MODULE LIMITATIONS

The C3k module is a key component of YOLO11, designed to enhance feature extraction capabilities through cross-stage feature reuse. Inspired by the C3 module in YOLOv5, it introduces a variable k to control the size of the convolutional kernel. Its basic structure consists of two parallel paths: a direct path and a densely connected path. As shown in Figure 4 below:

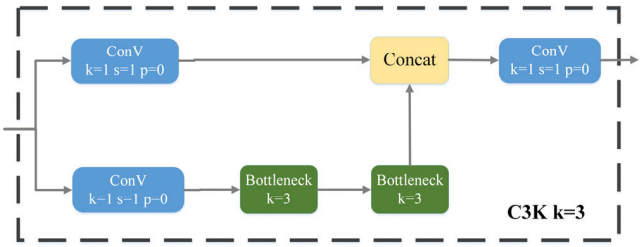

FIGURE 4. C3k module (when k=3).

Given an input feature map $\boldsymbol{X} \in \mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$, the processing of the C3k module can be described as follows: The input is first passed through two 1×1 convolutions, expressed as: $\boldsymbol{X_1} = \text{Conv}_{1\times1}(\boldsymbol{X}), \boldsymbol{X_2} = \text{Conv}_{1\times1}(\boldsymbol{X})$ Here, $\boldsymbol{X_1}$ follows the direct path: $\boldsymbol{Y_1} = \boldsymbol{X_1}$ While $\boldsymbol{X_2}$ goes through a densely connected path: $\boldsymbol{Y_2}$= Bottleneck(Bottleneck(Conv($\boldsymbol{X_2}$))).

Finally, the output is computed as: $\boldsymbol{Y} = \text{Conv}_{1\times1}(\text{Concat}(\boldsymbol{Y_1},\boldsymbol{Y_2}))$.

At the later stages of the Backbone, the focus should be on extracting multiple features rather than simply compressing them with convolutions. However, the traditional C3k module relies solely on conventional local convolution operations, assigning the same weights to all channel features. This approach fails to effectively distinguish and highlight critical features, and it lacks the ability to integrate the surrounding contextual information of license plates. Consequently, it struggles to capture the global information of vehicle body license plates.

This limitation in feature processing poses a bottleneck to further improving model accuracy. To address this, integrating attention mechanisms into the existing architecture offers a viable approach to enhance the feature extraction capability of the C3k module.

B. IMPROVEMENTS TO THE C3K2 MODULE

To address the limitations of the C3k module in feature extraction, this paper proposes a feature enhancement strategy based on an improved Multi-Head Self-Attention (MHSA). Through a systematic study of different attention mechanism integration schemes, six MHSA integration methods were designed and each method is described in detail below.

Serial Integration (Series-MHSA-C3K) is an intuitive feature enhancement strategy that integrates the MHSA module sequentially into the feature processing pipeline. As shown in Figure 5(a), based on the C3k architecture, the MHSA module is placed after the Bottleneck modules, enhancing feature extraction capability through sequential processing, followed by feature aggregation. This approach is straightforward but may increase the computational depth of the model.

Parallel Integration (Parallel-MHSA-C3K) [33] adopts a dual-path architecture for parallel feature processing. As shown in Figure 5(b), the input feature X2 is processed through two independent paths: one path uses the traditional Bottleneck structure for local feature extraction, while the other path employs MHSA to capture global context, ultimately achieving feature fusion. This design enables the model to simultaneously acquire local details and global semantic information.

Pyramid Integration (Pyramid-MHSA-C3K) [34] introduces a feature pyramid structure to enhance multi-scale feature representation. As shown in Figure 5(c), the original Bottleneck module is removed, constructing a dual-branch feature pyramid: one branch processes original-scale features, while the other processes downsampled features, implementing cross-scale feature aggregation through upsampling and feature concatenation. This design enhances the model's adaptability to targets of different scales.

Residual Integration (Residual-MHSA-C3K) [35] draws inspiration from ResNet's skip connection mechanism. As shown in Figure 5(d), hierarchical features are first extracted through a serial approach, followed by the introduction of residual connections with learnable weight ω for adaptive residual mapping. This design effectively preserves original feature information and mitigates the gradient vanishing problem in deep networks.

Gating Integration (Gating-MHSA-C3K) [36] introduces an adaptive gating mechanism for dynamic feature selection. As shown in Figure 5(e), the mathematical expression for this method is:

$$\begin{aligned}\mathbf{GW} &= \mathbf{Sigmoid}(\text{Conv}_{1\times1}(\text{MHSA}((\text{Conv}(\boldsymbol{X_2})))\\ \boldsymbol{Y_2} &= \mathbf{GW} \times \text{Bottleneck}(\text{Bottleneck}(\text{Conv}(\boldsymbol{X_2}))) \quad (4)\end{aligned}$$

where $\mathbf{GW}$ is a weight matrix of the same size as $\boldsymbol{X_2}$, which is an adaptive coefficient obtained through a series of transformations on the input feature $\boldsymbol{X_2}$; $\text{Conv}_{1\times1}$ denotes a 1×1 convolution operation used to adjust the feature channel

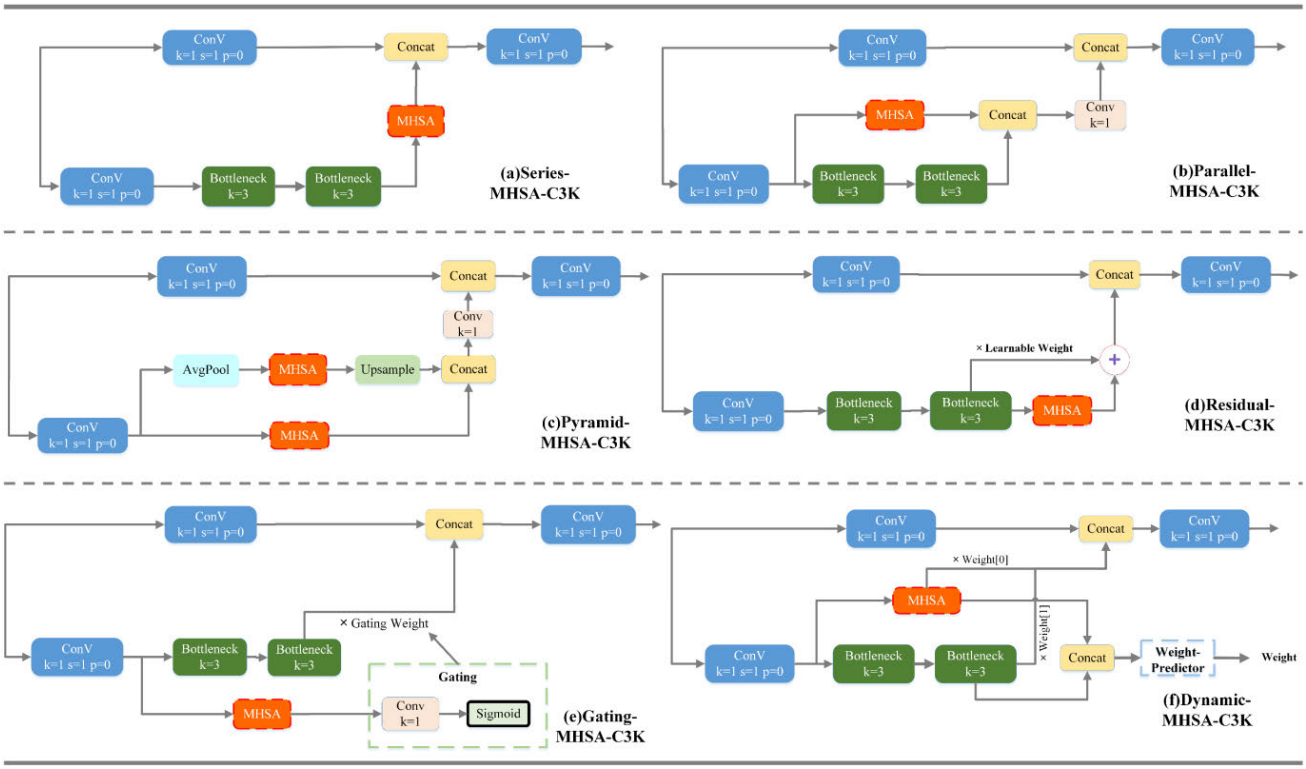

FIGURE 5. Six methods for improving the C3k module.

dimensions; MHSA represents the multi-head self-attention mechanism responsible for capturing global contextual information in the features; the Sigmoid function maps the output to the interval [0, 1], making it suitable as a weight coefficient; $\boldsymbol{Y_2}$ is the feature output enhanced by the gating mechanism, obtained through the element-wise product of the gating weight and Bottleneck features.

This gating connection approach is lightweight, requiring only three small modules to implement adaptive feature selection. This lightweight design enables the model to maintain computational efficiency while adaptively adjusting attention distribution according to the importance of input features, effectively enhancing feature representation in key regions while suppressing interference from irrelevant background noise.

Dynamic Weight Integration (Dynamic-MHSA-C3K) [37] employs a dedicated weight predictor for adaptive feature fusion. As shown in Figure 5(f), features are extracted through parallel paths: one path processed by MHSA and the other by Bottleneck, with fusion weights predicted by a lightweight weight predictor. Adopting a squeeze-and-excitation structure, it models the channel-wise feature importance, ultimately achieving dynamic feature aggregation.

C. GMH-YOLO OVERVIEW

An in-depth comparison of the six improvement strategies in the experimental section found that **Gating Integration (Gating-MHSA-C3K)** achieved the best balance between feature extraction capability and computational efficiency. Particularly when handling challenging targets such as vehicle bodies and license plates in construction site scenarios, its adaptive gating mechanism excels in selecting and enhancing effective features while suppressing background noise interference, resulting in more accurate license plate detection.

Based on this finding, the **GMH-YOLO (Gating Multi-Head Attention YOLO)** model is proposed by replacing the C3k in C3k2 with **Gating-MHSA-C3k**. This model integrates the gating attention mechanism into the C3k2 module. The overall framework of GMH-YOLO is shown in Figure 6.

GMH-YOLO retains the YOLO11 foundational architecture while incorporating a GMH-C3k into the Backbone. In the Backbone, the model adopts a progressive feature processing strategy characterized by "feature compression, preliminary extraction, attention enhancement, and multi-scale feature fusion." First, feature compression is effectively achieved through the Conv layer. Then, the C3K2F module is used for preliminary feature extraction, enhancing the foundational representational capacity of the features.

In the middle and later stages of the network, two GMH-C3k modules (marked in red in Figure 6) are strategically placed to adaptively enhance attention to features through the gating mechanism. This significantly improves the model's ability to recognize challenging targets such as vehicle bodies and license plates. Finally, multi-scale feature fusion and spatial attention optimization are achieved through the SPPF,

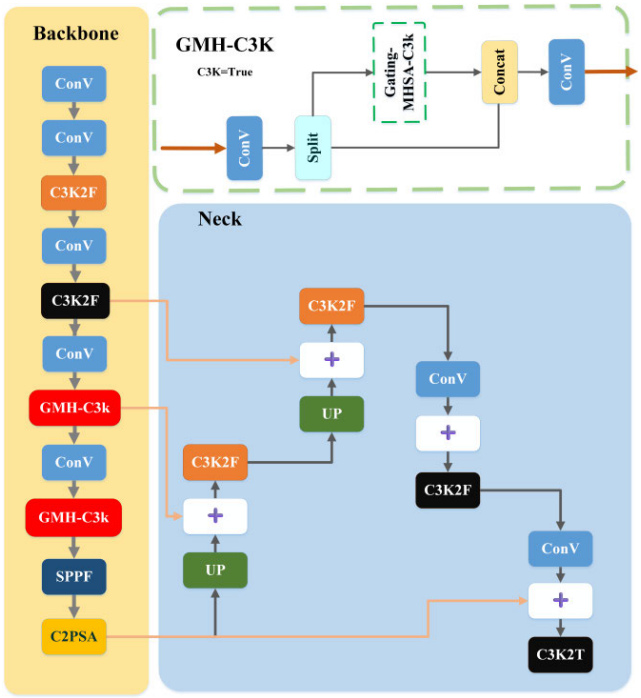

FIGURE 6. GMH-YOLO framework diagram(backbone and neck), red modules indicate added parts, black modules indicate where the test is added.

C2PSA modules, and PANet architecture, providing rich feature representations for detection tasks.

To validate the effectiveness of module placement, comparative experiments were also conducted by placing the modules in other positions (as indicated by the black modules in the diagram). The experimental results demonstrate that adding two GMH-C3k modules in the latter half of the Backbone is the optimal configuration. This arrangement ensures sufficient and appropriate feature extraction while maintaining computational efficiency.

IV. EXPERIMENTAL

A. EXPERIMENT PLATFORM AND EVALUATION METRICS

In this study, experiments were conducted on a system equipped with an Intel Core i7-12800H eight-core processor (base frequency 2.30 GHz, maximum turbo frequency 4.60 GHz) and an NVIDIA GeForce RTX 3060 GPU for training and testing. For the software environment, GMH-YOLO was implemented using Python 3.10.15 and the CUDA-enabled PyTorch 2.5.1 framework. During the training phase, the model was trained on the previously segmented CSLPD dataset and employed the pre-trained weights of YOLO11n for transfer learning. The training process utilized a multi-scale training strategy with a batch size of 8. The initial learning rate was set to 0.01 and adjusted using a cosine annealing scheduler. The optimizer was set to SGD, the random seed was fixed at 0, and the model was trained for 300 epochs to ensure adequate learning of the dataset features and to achieve optimal performance.

For evaluation metrics, given that the detected anchor boxes do not require high precision as long as they fully encompass the license plates, the detection accuracy for license plates was assessed using **mAP50**, with **mAP50-90** as a supplementary metric. To evaluate the model's computational complexity and real-time capability, metrics such as **the number of model parameters, GFLOPS, and GPS** [38] were employed.

B. DATASET

Construction site vehicle recognition presents distinct challenges due to environmental complexities such as dust, uneven lighting, and wet surfaces. Construction vehicles typically feature multiple identification markers: standard license plates, auxiliary plates indicating vehicle type (e.g., dump trucks), and license information sprayed on vehicle bodies for multi-angle visibility.

Existing license plate datasets, such as the Real-time dataset [39] for highway monitoring and CCPD [40] for parking management, do not adequately address construction site-specific requirements. These datasets lack the unique characteristics found in construction environments.

To bridge this gap, we developed the Construction Site License Plate Dataset (CSLPD). Data collection was conducted across three representative construction site types (building, road, and hydraulic engineering). We utilized 720p HD cameras (1280 × 720) deployed at critical license plate detection areas such as site entrances and main vehicle passages. To capture multi-angle views, cameras were installed at various heights (approximately 1.5, 3, and 5 meters) and covered multiple shooting distances (5-30 meters) to ensure data diversity. We captured images under various conditions, including different time periods and weather types (clear, rainy, and hazy). Special attention was paid to construction site-specific factors like dust and concrete powder that affect license plate recognition. The collected vehicle types comprehensively covered common construction vehicles, including concrete mixers, dump trucks, and other engineering vehicles.

Data collection continued for approximately 20 days, utilizing a combination of continuous video recording and frame extraction at 60-frame intervals, initially obtaining over 5000 valid images. Considering the redundancy of similar scene images, we established a filtering rule of "preserving a maximum of 10 images for similar locations and vehicle types," and through rigorous manual screening, ultimately retained 1495 high-quality and challenging images. Figure 7 shows examples of these images, illustrating the diversity and complexity of our dataset. Professional annotators used the LabelImg tool to precisely label yellow license plates, blue license plates, and body-painted license plates following the VOC dataset standard. We extended the standard annotation format to include additional attributes for challenging conditions (as shown on the right side of Figure 8, including high-speed, dust, strong light conditions and so

FIGURE 7. Example of CSLPD dataset (red box indicates body license plate, orange box indicates high speed, yellow box indicates dust impact, blue box indicates strong light, and purple box indicates weak light).

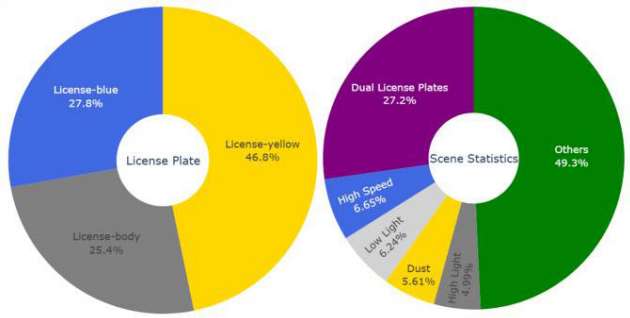

FIGURE 8. The left figure shows the category statistics of license plates in CSLPD, and the right figure shows the statistics of complex environments.

on), which allows us to better analyze and understand the predicted results we obtain later. This multi-dimensional annotation approach resulted in a comprehensive dataset containing 2,301 license plate instances with rich contextual information.

Given the relatively small size of dataset, the CSLPD dataset was randomly divided into training, validation, and test sets in a 7:1:2 ratio. Specifically, the training set contains 1,046 images, the validation set contains 150 images, and the test set contains 299 images. During the division process, we ensured that the distribution of different scene types (e.g., dual license plates, body license plates) was consistent across the subsets to ensure the representativeness of the evaluation results.

C. EXPERIMENTAL RESULTS

In our experimental approach, we carefully considered both detection performance and computational efficiency. Current license plate recognition models (such as YOLO-SLD, LPR-Net) typically have a large number of parameters (as shown in Table 4) and high GFLOPs, requiring costly computational resources. Therefore, we adopted the lightweight YOLO11 as our foundational model, while referencing improvement approaches from specialized license plate recognition models as our baseline for comparison. Following the strategy used in YOLO-SLD, we integrated CBAM and SimAM into the C3k2 module for comparative experiments. The results are shown in Table 1.

Compared to YOLO11, integrating different attention mechanisms led to an increase in model parameters across the board. Although the proposed GMH-YOLO increased the parameter count by 4.2%, the optimized gating mechanism reduced GFLOPs by 0.1, demonstrating excellent computational efficiency. In terms of real-time performance, GMH-YOLO's FPS decreased only slightly by 2.5% from YOLO11's 23.96 FPS to 23.36 FPS. Considering that vehicles at construction sites move at limited speeds and follow relatively fixed routes, this minimal frame rate difference has negligible impact on actual monitoring effectiveness, fully meeting the real-time monitoring requirements for construction site vehicle management. Notably, while YOLO11-SimAM's FPS only decreased from YOLO11's 23.96 to 23.8, its computational cost increased significantly with GFLOPs rising from 6.4 to 6.7, and it showed markedly lower accuracy in the critical vehicle body license plate detection task. In contrast, GMH-YOLO achieves the highest overall detection accuracy (93.3% mAP@50) while maintaining ideal computational efficiency by reducing GFLOPs, offering significant application advantages in computationally constrained construction site intelligent monitoring systems that require high accuracy.

A deeper analysis of performance metrics reveals that YOLO11's subpar performance in overall mAP@50 is primarily due to its significant deficiency in detecting body license plates, with related metrics at only 81.3%. In contrast, GMH-YOLO achieved a remarkable breakthrough in this challenging task, increasing the mAP@50 for body license plate detection to 87.2%, a substantial improvement of 5.9%, thereby driving a 1.4% increase in overall detection accuracy. While the models improved with CBAM and SimAM performed well in detecting standard license plates, they exhibited significant performance degradation in the critical task of body license plate detection. Notably, in the CBAM experiment, although the mAP@50 for blue license plates improved slightly by 0.9%, the detection performance for body license plates dropped sharply by 21.3%. These results strongly validate the theoretical analysis in Section II regarding the adaptability of different attention mechanisms in complex scenarios.

D. CONTRASTING EXPERIMENT

To systematically evaluate the effectiveness of the different improvement strategies proposed in Section III-B, we conduct comparative experiments on six attention mechanism integration strategies (as shown in Table 2). The experimental results demonstrate that GMH-YOLO not only maintains high accuracy in standard license plate detection but also achieves a significant breakthrough in the challenging task of body license plate recognition.

We further investigated the impact of module configuration on detection performance by designing a series of comparative experiments, with results shown in Table 5. When

TABLE 1. Baseline model comparison.

Method	Para	GFLOPs	FPS		L-Yellow	L-Blue	L-Body	All
YOLO8	3.01M	12.8	23.02	mAP_{50}	0.995	0.947	0.69	0.877
				$mAP_{50:95}$	0.687	0.664	0.449	0.6
YOLO11	2.59M	6.4	23.96	mAP_{50}	0.995	0.949	0.813	0.919
				$mAP_{50:95}$	0.678	0.638	0.519	0.612
YOLO11-CBAM	2.98M	7.8	23.02	mAP_{50}	0.994	0.958	0.6	0.851
				$mAP_{50:95}$	0.677	0.594	0.318	0.53
YOLO11-SimAM	2.73M	6.7	23.8	mAP_{50}	0.995	0.962	0.754	0.903
				$mAP_{50:95}$	0.66	0.638	0.407	0.568
Proposed	**2.70M**	**6.3**	**23.36**	mAP_{50}	**0.995**	**0.931**	**0.872**	**0.933**
				$mAP_{50:95}$	**0.693**	**0.636**	**0.515**	**0.615**

TABLE 2. Comparison of 6 improvement methods in MHSA.

Method	Para	GFLOPs	FPS		L-yellow	L-blue	L-body	All
YOLOSer-MHSA	2.59M	6.3	23.92	mAP_{50}	0.993	0.939	0.712	0.881
				$mAP_{50:95}$	0.674	0.639	0.459	0.591
YOLOPar-MHSA	2.71M	6.2	22.65	mAP_{50}	0.994	0.934	0.607	0.845
				$mAP_{50:95}$	0.666	0.621	0.42	0.569
YOLOPyr-MHSA	2.72M	6.3	21.62	mAP_{50}	0.995	0.94	0.71	0.882
				$mAP_{50:95}$	0.671	0.614	0.469	0.584
YOLORes-MHSA	2.70M	6.3	22.3	mAP_{50}	0.995	0.944	0.677	0.872
				$mAP_{50:95}$	0.656	0.616	0.438	0.57
YOLODyn-MHSA	2.70M	6.3	23.97	mAP_{50}	0.992	0.954	0.716	0.887
				$mAP_{50:95}$	0.643	0.599	0.444	0.562
Proposed	**2.70M**	**6.3**	**23.36**	mAP_{50}	**0.995**	**0.931**	**0.872**	**0.933**
				$mAP_{50:95}$	**0.693**	**0.636**	**0.515**	**0.615**

TABLE 3. Ablation experiment of YOLO11.

Method	Para	GFLOPs	mAP_{50}	$mAP_{50:95}$
Yolo11	2.59M	6.4	0.919	0.621
Yolo11-Without-C2PSA	2.33M	6.1	0.913	0.603
Yolo11-Without-C3k2	2.50M	6.2	0.858	0.563
Yolo11-Without-C3k2&C2PSA	2.26M	6.0	0.843	0.539

TABLE 4. Test on CCPD Dataset.

Method	Para	FPS	mAP_{50}	$mAP_{50:95}$
YOLO-SLD	70.1M	21.49	0.993	0.904
YOLO11	2.58M	34.71	0.995	0.919
YOLO-SimAM	2.69M	35.29	0.995	0.918
Proposed	**2.70M**	**34.63**	**0.995**	**0.921**

the last two C3k modules in the Backbone were replaced with GMH-C3k, the model achieved optimal performance. This result has inherent logic: the latter part of the Backbone is responsible for extracting higher-level semantic features. Deploying Gating-MHSA at this stage can more effectively integrate global contextual information while maintaining sensitivity to local features through the gating mechanism.

To explore the effectiveness of the gating mechanism, the configuration of two modules at the end of the Backbone was retained, and experiments were conducted using the same six improvement strategies with SimAM (as

TABLE 5. Exploration of increasing the number of different modules.

Method	Para	GFLOPs	FPS		L-yellow	L-blue	L-body	All
GMH-YOLO-1B	2.67M	6.3	23.96	mAP_{50}	0.994	0.96	0.763	0.906
				$mAP_{50:95}$	0.67	0.658	0.484	0.604
GMH-YOLO-2B	**2.70M**	**6.3**	**23.36**	mAP_{50}	**0.995**	**0.931**	**0.872**	**0.933**
				$mAP_{50:95}$	**0.693**	**0.636**	**0.515**	**0.615**
GMH-YOLO-3B	2.71M	6.4	20.86	mAP_{50}	0.979	0.929	0.877	0.928
				$mAP_{50:95}$	0.674	0.619	0.523	0.605
GMH-YOLO-1H	2.80M	6.4	21.62	mAP_{50}	0.995	0.966	0.727	0.896
				$mAP_{50:95}$	0.662	0.638	0.421	0.574
GMH-YOLO-2H	2.83M	6.4	22.75	mAP_{50}	0.985	0.917	0.661	0.855
				$mAP_{50:95}$	0.68	0.615	0.408	0.568

TABLE 6. Comparison of 6 improvement methods in SimAM.

Method	Para	GFLOPs	FPS		L-yellow	L-blue	L-body	All
YOLOSer-SimAM	2.59M	6.4	23.73	mAP_{50}	0.995	0.886	0.768	0.883
				$mAP_{50:95}$	0.69	0.597	0.466	0.584
YOLOPar-SimAM	2.69M	6.3	23.2	mAP_{50}	0.995	0.933	0.832	0.92
				$mAP_{50:95}$	0.65	0.621	0.518	0.596
YOLOPyr-SimAM	2.69M	6.1	23.25	mAP_{50}	0.995	0.931	0.785	0.904
				$mAP_{50:95}$	0.676	0.616	0.46	0.585
YOLORes-SimAM	2.68M	6.3	23.37	mAP_{50}	0.995	0.96	0.827	0.923
				$mAP_{50:95}$	0.656	0.66	0.502	0.606
YOLOGAI-SimAM	**2.69M**	**6.3**	**24.41**	$\boldsymbol{mAP_{50}}$	**0.995**	**0.942**	**0.837**	**0.925**
				$\boldsymbol{mAP_{50:95}}$	**0.676**	**0.653**	**0.511**	**0.613**
YOLODyn-SimAM	2.68M	6.3	23.0	mAP_{50}	0.994	0.945	0.739	0.893
				$mAP_{50:95}$	0.66	0.65	0.416	0.576

shown in Table 6). The experimental results show that although accuracy decreased compared to GMH-YOLO, the improved Gating-SimAM achieved a 0.6% mAP@50 improvement over the original YOLO11. Moreover, Gating-SimAM demonstrated more significant advantages in real-time performance.

E. ABLATION EXPERIMENT

To more deeply analyze the contribution of key modules in the YOLO11 model, we conducted ablation experiments. For the C2PSA module, we chose to remove it directly. As for the C3k2 module, we reverted it to the YOLOv8 version, replacing it with the C2f module for alternative processing. The ablation results are shown in Table 3.

The experimental results demonstrate that both C2PSA and C3k2 modules significantly impact model performance. Removing the C2PSA module led to a 0.6% decrease in mAP@50, indicating that this module effectively enhances detection precision through its enhanced spatial attention mechanism. Removal of the C3k2 module resulted in a more substantial performance degradation (6.1% decrease in mAP@50), validating its critical role in feature extraction. When both modules were removed simultaneously, the mAP@50 declined by 7.6%, confirming the importance of these structural innovations to the YOLO11 framework.

F. CONFIRMATORY EXPERIMENT ON PUBLIC DATASET

To validate the generalizability and transferability of the GMH-YOLO model, comparative experiments were conducted on the public license plate dataset CCPD [40]. Due to the current lack of public datasets specifically targeting construction site license plate recognition, we selected the widely-used single license plate dataset CCPD for testing, randomly sampling 20,000 images as test samples. The results are shown in Table 4.

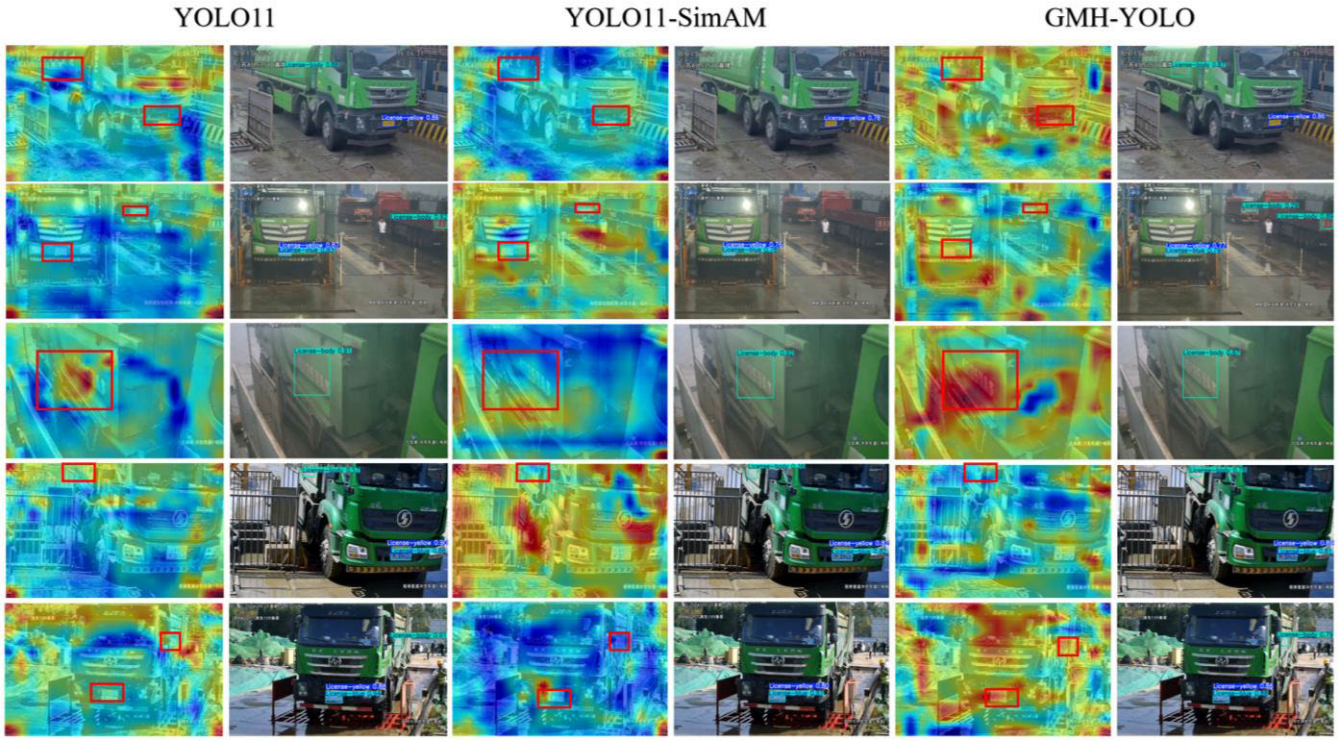

FIGURE 9. YOLO11, YOLO11-SimAM and GMH-YOLO heatmap.

Experimental results demonstrate that YOLO11 series models significantly improve computational efficiency while reducing parameter count compared to the recently proposed YOLO-SLD, while maintaining high detection accuracy. Although our proposed GMH-YOLO model shows a slight decrease in FPS compared to YOLO11 (approximately 0.08), it achieves 92.1% mAP@[50:95], outperforming the baseline YOLO11 model by 0.2%. These results verify that our proposed attention mechanism not only excels in complex construction site scenarios but also remains effective in general license plate detection tasks, demonstrating excellent cross-scenario adaptability.

G. VISUAL COMPARISON

To more intuitively demonstrate the improvements of GMH-YOLO, Figure 9 presents attention heatmaps obtained by visualizing feature maps from the detection head and alpha-blending them with original images, comparing YOLO11, YOLO11-SimAM, and GMH-YOLO. The heatmap visualization clearly shows the precision of attention focusing, with red areas representing high attention regions and blue areas representing low attention regions.

Specifically, as shown in the third row, when processing body license plates, GMH-YOLO demonstrates more accurate and concentrated attention distribution compared to YOLO11 and YOLO11-SimAM. In scenarios containing both standard license plates and body license plates (as shown in rows 1, 2, 4, and 5), GMH-YOLO maintains a high level of attention to license plates of varying sizes. Its heatmaps exhibit more balanced feature responses, as demonstrated by the detection results in the second row, where the model successfully detects distant body license plates that YOLO11 and YOLO11-SimAM failed to recognize. Furthermore, the heatmaps reveal that GMH-YOLO shows clearer feature responses in license plate boundary areas. This enhanced boundary-awareness improves the positioning accuracy of detection boxes, especially for body license plates with less distinct boundaries.

Overall, the red highlighted areas in GMH-YOLO's heatmaps closely correspond to license plate positions, validating the effectiveness of its attention mechanism. Compared to YOLO11 and YOLO11-SimAM, GMH-YOLO's attention distribution is more focused, enabling the model to better handle complex license plate recognition tasks in construction site environments, particularly excelling in detecting challenging body license plates. These visualization results provide an intuitive explanation for GMH-YOLO's 5.9% performance improvement in body license plate detection tasks, confirming the effectiveness of the gated multi-head attention mechanism in enhancing the model's feature extraction capability.

H. EXPERIMENT RESULTS

Based on a comprehensive experimental analysis, GMH-YOLO demonstrated significant performance advantages in construction site license plate detection tasks. In terms

of detection accuracy, the model achieved an mAP@50 of 93.3% on the CSLPD dataset, a 1.4% improvement over the baseline model YOLO11. Notably, while maintaining high recognition rates for standard license plates (yellow plates: 99.5%, blue plates: 93.1%), GMH-YOLO made breakthrough progress in the challenging body license plate detection task, increasing the mAP@50 from 81.3% to 87.2%, a 5.9% improvement. Heatmap visualizations further validated the model's precise localization of key features in license plate regions.

In terms of computational efficiency, the optimized gating mechanism design allowed GMH-YOLO to achieve a good balance between performance improvement and computational cost. Specifically, the model's parameter count increased by only 4.2%, while GFLOPs decreased from 6.4 to 6.3. Although FPS dropped slightly (from 23.96 to 23.36), it remains fully sufficient for real-time detection requirements. Compared to existing mainstream attention mechanisms for license plate, GMH-YOLO demonstrated unique advantages in complex construction site scenarios. For example, compared to CBAM, GMH-YOLO avoided the significant drop in body license plate detection performance observed when improving standard license plate recognition rates; compared to SimAM, it achieved more substantial performance improvements in the critical task of body license plate detection while preserving a lightweight model architecture.

Additionally, through theoretical analysis and experimental validation of various improvement strategies, this study confirmed the effectiveness of the gating attention mechanism, providing valuable technical references and improvement ideas for future research in related fields. The experimental results show that GMH-YOLO not only achieves a balanced performance across multiple metrics but also makes significant breakthroughs in the most challenging real-world application of body license plate detection.

V. CONCLUSION

This study addresses the specific challenges of license plate detection in construction site environments by constructing the CSLPD dataset and designing the GMH-YOLO model, providing an efficient and reliable solution for construction site license plate recognition. In terms of dataset development, we constructed the CSLPD dataset specifically for construction site scenarios, containing 1,495 images and 2,301 license plate instances. These data cover complex scenes such as double license plates and body license plates, as well as environmental factors unique to construction sites such as strong lighting and dust interference, providing a valuable data resource for related research.

We analyzed the limitations of the C3k module in processing complex scenes from both theoretical and practical perspectives, and proposed six different attention mechanism integration strategies. Comparative experiments demonstrated that the gating integration method achieved the optimal balance between feature extraction capability and computational efficiency. Based on this finding, we designed the GMH-YOLO model, integrating a gated multi-head attention mechanism into the C3k2 module. Experimental results verified the effectiveness of GMH-YOLO, achieving 93.3% mAP@50 on the CSLPD dataset, an improvement of 1.4% over the baseline model. While this improvement may seem modest, it has significant implications in practical applications: for construction sites with 200 vehicles entering and exiting daily, a 1.4% accuracy improvement means approximately 3 fewer misidentifications per day, significantly reducing the need for manual intervention in vehicle registration and minimizing traffic delays.

Notably, in the most challenging task of body license plate detection, accuracy increased from 81.3% to 87.2%, a substantial improvement of 5.9%. This is particularly important for complex construction site scenarios, especially when mud contamination or strong light reflection makes standard license plates difficult to recognize, where high-precision detection of body license plates provides a crucial backup identification pathway.

Furthermore, in terms of model deployment, GMH-YOLO maintained excellent real-time performance while improving detection accuracy. Despite a 4.2% increase in parameter count to 2.70M, the innovative gating mechanism design actually reduced computational complexity, decreasing GFLOPs from 6.4 to 6.3. In terms of inference speed, FPS only decreased slightly from 23.96 to 23.36, less than a 3% reduction. This computational efficiency enables GMH-YOLO to run effectively on edge computing devices. Its low power consumption characteristics make it suitable for deployment in resource-constrained construction site environments.

Despite these significant improvements, our model still faces certain limitations. While the accuracy for body license plates increased substantially to 87.2%, it still falls short of application-level precision requirements for industrial deployment, remaining a challenging area that requires further research attention. Additionally, although the current CSLPD dataset captures complex scenes specific to construction sites, its sample size is relatively limited and primarily focused on daytime conditions with good lighting. The model's performance in extreme weather conditions or under complex lighting scenarios has yet to be thoroughly evaluated. These limitations highlight the need for future research to expand the dataset's scale and diversity, particularly by incorporating samples under various environmental challenges such as heavy snow, fog, rain, and nighttime conditions.

In the future, we will focus on addressing the current limitations of the model. We plan to expand the scale and diversity of the CSLPD dataset by adding samples under extreme weather conditions such as heavy snow, fog, and rain to enhance the model's robustness in adverse environments. Meanwhile, we will explore more lightweight architectural designs and specialized attention mechanisms aimed at further improving detection accuracy. These research directions

will further unlock the application potential of license plate recognition technology in construction site management and promote the digital transformation of the construction industry.

ACKNOWLEDGMENT

The authors gratefully acknowledge Ultralytics for providing the open-source implementation of YOLO, which has been instrumental in this work.

REFERENCES

[1] S. Paneru and I. Jeelani, "Computer vision applications in construction: Current state, opportunities & challenges," *Autom. Construct.*, vol. 132, Dec. 2021, Art. no. 103940.

[2] J. R. Montoya-Torres, S. Moreno, W. J. Guerrero, and G. Mejía, "Big data analytics and intelligent transportation systems," *IFAC-PapersOnline*, vol. 54, no. 2, pp. 216–220, Jan. 2021.

[3] H. Arab, I. Ghaffari, L. Chioukh, S. O. Tatu, and S. Dufour, "A convolutional neural network for human motion recognition and classification using a millimeter-wave Doppler radar," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4494–4502, Mar. 2022.

[4] Y. Yuan, W. Zou, Y. Zhao, X. Wang, X. Hu, and N. Komodakis, "A robust and efficient approach to license plate detection," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1102–1114, Mar. 2017.

[5] C.-Y. Lei and J.-H. Liu, "Vehicle license plate character segmentation method based on watershed algorithm," in *Proc. Int. Conf. Mach. Vis. Human-Mach. Interface*, Apr. 2010, pp. 447–452.

[6] L. Galarza, H. Martin, and M. Adjouadi, "Time-of-Flight sensor in a book reader system design for persons with visual impairment and blindness," *IEEE Sensors J.*, vol. 18, no. 18, pp. 7697–7707, Sep. 2018.

[7] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extraction*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023.

[8] S. Luo and J. Liu, "Research on car license plate recognition based on improved YOLOv5m and LPRNet," *IEEE Access*, vol. 10, pp. 93692–93700, 2022.

[9] M.-A. Chung, Y.-J. Lin, and C.-W. Lin, "YOLO-SLD: An attention mechanism-improved YOLO for license plate detection," *IEEE Access*, vol. 12, pp. 89035–89045, 2024.

[10] Y. Quan, P. Wang, Y. Wang, and X. Jin, "GUI-based YOLOv8 license plate detection system design," in *Proc. 5th Int. Conf. Control Robot. (ICCR)*, Nov. 2023, pp. 156–161.

[11] R. Khanam and M. Hussain, "YOLOv11: An overview of the key architectural enhancements," 2024, *arXiv:2410.17725*.

[12] P. Tarabek, "A real-time license plate localization method based on vertical edge analysis," in *Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Sep. 2012, pp. 149–154.

[13] J.-W. Hsieh, S.-H. Yu, and Y.-S. Chen, "Morphology-based license plate detection from complex scenes," in *Proc. Int. Conf. Pattern Recognit.*, Aug. 2002, pp. 176–179.

[14] S.-L. Chang, L.-S. Chen, Y.-C. Chung, and S.-W. Chen, "Automatic license plate recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 5, no. 1, pp. 42–53, Mar. 2004.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Proc. ECCV*, vol. 9905, Cham, Switzerland, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Springer, 2016, pp. 21–37.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[18] V. Agarwal and G. Bansal, "Automatic number plate detection and recognition using YOLO world," *Comput. Electr. Eng.*, vol. 120, Dec. 2024, Art. no. 109646.

[19] H. Moussaoui, N. E. Akkad, M. Benslimane, W. El-Shafai, A. Baihan, C. Hewage, and R. S. Rathore, "Enhancing automated vehicle identification by integrating YOLO v8 and OCR techniques for high-precision license plate detection and recognition," *Sci. Rep.*, vol. 14, no. 1, p. 14389, Jun. 2024, doi: 10.1038/s41598-024-65272-1.

[20] A. Ismail, M. Mehri, A. Sahbani, and N. Essoukri Ben Amara, "A dual-stage system for real-time license plate detection and recognition on mobile security robots," *Robotica*, vol. 2025, pp. 1–22, Jan. 2025, doi: 10.1017/s0263574724001991.

[21] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 1571–1580.

[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[23] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, and J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13728–13737.

[24] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," 2024, *arXiv:2405.14458*.

[25] C. Chen, X. Sun, H. Yang, J. Dong, and H. Xv, "Learning deep relations to promote saliency detection," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, pp. 10510–10517, doi: 10.1609/aaai.v34i07.6622.

[26] X. Sun, C. Chen, X. Wang, J. Dong, H. Zhou, and S. Chen, "Gaussian dynamic convolution for efficient single-image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2937–2948, May 2022, doi: 10.1109/TCSVT.2021.3096814.

[27] L. Yang, R. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 11863–11874.

[28] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "FFCA-YOLO for small object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5611215, doi: 10.1109/TGRS.2024.3363057.

[29] Y. Cai, Z. Yao, H. Jiang, W. Qin, J. Xiao, X. Huang, J. Pan, and H. Feng, "Rapid detection of fish with SVC symptoms based on machine vision combined with a NAM-YOLO v7 hybrid model," *Aquaculture*, vol. 582, Mar. 2024, Art. no. 740558.

[30] C. Xue, Y. Xia, M. Wu, Z. Chen, F. Cheng, and L. Yun, "EL-YOLO: An efficient and lightweight low-altitude aerial objects detector for onboard applications," *Expert Syst. Appl.*, vol. 256, Dec. 2024, Art. no. 124848.

[31] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, "ViT-YOLO: Transformer-based YOLO for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2799–2808.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, Jun. 2017, pp. 5998–6008.

[33] M. Zhu and Y. Wu, "A parallel convolutional neural network for pedestrian detection," *Electronics*, vol. 9, no. 9, p. 1478, Sep. 2020.

[34] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[35] R. Tian, Z. Wu, Q. Dai, H. Hu, Y. Qiao, and Y.-G. Jiang, "ResFormer: Scaling ViTs with multi-resolution training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22721–22731.

[36] X. Liang and C. Jung, "AGNet: Attention guided sparse depth completion using convolutional neural networks," *IEEE Access*, vol. 10, pp. 10514–10522, 2022.

[37] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C. Hsieh, "DynamicViT: Efficient vision transformers with dynamic token sparsification," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Jan. 2021, pp. 13937–13949.

[38] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, and B. Lai, "PP-YOLOE: An evolved version of YOLO," 2022, *arXiv:2203.16250*.
[39] A. Agarwal, A. Thombre, K. Kedia, and I. Ghosh, "ITD: Indian traffic dataset for intelligent transportation systems," in *Proc. 16th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2024, pp. 842–850.
[40] H. Li, P. Wang, and C. Shen, "Toward end-to-end car license plate detection and recognition with deep neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1126–1136, Mar. 2019.

MING LI received the Ph.D. degree from Nanjing University of Science and Technology.

He is currently an Associate Professor with the Business School, Hohai University. He leads the National Social Science Fund project on engineering construction market supervision. He has published in internationally renowned journals, including *Managerial and Decision Economics*. His research covers innovative topics, such as dynamic pricing strategies and intelligent construction of water conservancy projects using BIM+IoT technology. His research focuses on information systems and engineering management, particularly in engineering project management informatization and system dynamics modeling.

ZE-QUAN WANG is currently pursuing the degree in information management and information systems with Hohai University. His current research focuses on object detection and tracking, as well as applications involving UE5-based drone simulation and large language models. His research interests include autonomous driving, deep learning, multimodal systems, mobile applications, and artificial intelligence.

YU-HANG ZHAO is currently pursuing the degree in information management and information systems with Hohai University. His main research interests include object detection, multimodal systems, and deep learning.

QIANG LI received the bachelor's degree in electrical engineering and automation from Southeast University, in 2008, and the master's degree in project management from Hohai University, in 2021. He is currently a Senior Project Manager and an Electronic Information Engineer with Zhongbo Information Technology Research Institute Company Ltd. He has been engaged in informatization management of government and enterprise projects for a long time. He has in-depth research on safety supervision informatization of special equipment. His research interests include wireless communication, intelligent robots, autonomous vehicle, AI, and big data algorithm research.