Journal of Beijing Institute of Technology, 2026, Vol. 35, No. 2

YOLOv8-RGBT-MT: A Fast Multi-Modal Fusion Network for Worker-and-Equipment Detection on Construction Sites

Yan Li, Cunxin Sun, Baihai Zhang✉
(School of Automation, Beijing Institute of Technology, Beijing 100081, China)

Abstract: To address the challenges of dusty, foggy and other complex construction site environments leading to the failure of visible light imaging and difficulties in small target detection, as well as the high resource consumption hindering model deployment, an enhanced and lightweight algorithm is proposed. This algorithm employs a hybrid architecture, integrating red green blue (RGB) (visible light) and thermal infrared (RGBT) multi-modal images through a fusion framework based on you only look once (YOLO) version 8 and Mamba-Transformer (MT). We refer to this integrated model as YOLOv8-RGBT-MT. In terms of network improvements, a frequency enhancement module is first employed to enhance visible light and infrared images. And then, a module integrating Mamba and Transformer components is designed to replace base convolutional blocks in the backbone network, thereby expanding the receptive field of the model and improving feature extraction in complex backgrounds. Finally, a multi-modal feature fusion mechanism is introduced, through which complementary information from visible and infrared images is effectively integrated via an adaptive weighting strategy, so that both the detection accuracy and robustness for small targets are enhanced. Experimental results demonstrate that, compared to YOLOv8-RGBT, the enhanced algorithm achieves an improvement of 18.7 % in mAP50, while reducing the number of inference time by 79.7 %

Keywords: RGBT images; multi-modal feature fusion; YOLOv8; Mamba-Transformer; real-time object detection

1 Introduction

With the rapid technological advancement, substantial progress has been made in the development of smart construction sites. The construction industry operates in dynamic and hazardous environments, where constantly moving entities such as personnel, materials, and machinery expose workers to numerous safety risks. In this context, smart construction sites play a vital role in enhancing onsite safety and optimizing resource allocation [1]. As the adoption of such intelligent systems continues to grow, construction site detection, a core component of smart construction environments, has attracted increasing attention from researchers [2–6].

However, the complex backgrounds of construction sites frequently result in the occlusion and misidentification of small targets, while airborne obscurants such as dust and fog can cause severe degradation in visible light images. Moreover, the limited computational resources available in such environments impose significant constraints on model complexity, making algorithms impractical for real-world deployment. These challenges collectively constitute critical barriers

Manuscript received Oct. 13, 2025; revised Nov. 19, 2025; accepted Nov. 26, 2025. The associate editor coordinating the review of this manuscript was Dr. Zhun Fan.
✉ Corresponding author. Email: smczhang@bit.edu.cn
DOI: 10.15918/j.jbit1004-0579.2025.079

to robust and efficient object detection in construction sites.

With the widespread adoption of deep learning in object detection, detecting objects on construction sites has emerged as a prominent research focus. Current object detection algorithms are broadly categorized into two-stage and single-stage approaches. Two-stage detectors, such as region-based convolutional neural networks (R-CNN) and fully convolutional networks (FCN), typically offer high accuracy but at the cost of significant computational overhead and limited real-time performance. In contrast, single-stage detectors provide faster inference speeds, making them more suitable for time-critical applications in dynamic construction environments. For instance, faster R-CNN has been adapted for safety helmet detection, demonstrating strong practical utility and effectiveness [7]. However, its performance tends to degrade under challenging environmental conditions, such as significant illumination variations. To address occlusion-related challenges, Espinosa-Oviedo et al. [8] modified the convolutional feature extraction component of faster R-CNN, while Ren et al. [9] integrated attention mechanisms to enhance detection accuracy in occluded scenes. Despite these improvements, the computational burden remains a limiting factor. More recently, Huang et al. [10] incorporated deformable convolutions into you only look once (YOLO) version 5, enabling more flexible feature perception and improving model robustness. Guo et al. [11] introduced a multi-scale detection model that leverages feature fusion and multi-scale training to enhance small object detection by capturing finer edge details. Similarly, Ge et al. [12] improved YOLO by replacing the faster CSP Bottleneck with 2 convolutions (C2f) module with an enhanced perceptive field (EPF) module to expand the receptive field, suppress background noise, and enhance feature extraction.

Nevertheless, deploying these advanced models in resource-constrained construction environments remains a significant challenge. To address this, Sun et al. [13] conducted a comprehensive evaluation of mainstream object detection algorithms and selected YOLOv7-tiny as a lightweight yet effective solution. In recent years, integrating YOLO-based frameworks with Transformer or State Space Models has gained considerable attention within the research community. These hybrid architectures aim to balance detection accuracy with computational efficiency, offering promising directions for real-time object detection in complex and resource-limited construction scenarios [14–22].

Building upon the aforementioned methods, this paper proposes an enhanced and lightweight detection algorithm, you only look once version 8 RGB and thermal infrared multi-modal Mamba-Transformer model (YOLOv8-RGBT-MT). The primary contributions of the algorithm are as follows:

1) The frequency enhancement module (FEM) is employed to enhance and denoise visible and infrared images, particularly under adverse weather conditions such as dust, fog, rain and snow.

2) A C2f module integrating Mamba and transformer block (C2f-MTB) is designed to replace the specific C2f module in the backbone network, thereby expanding the receptive field of the model and improving feature extraction in complex backgrounds.

3) A multi-modal feature fusion (MFF) mechanism is introduced to effectively integrate complementary information from visible and infrared images through an adaptive weighting strategy, enhancing both detection accuracy and robustness for small targets.

2 Method

2.1 Algorithm Framework

To address the challenges of complex construction site environments leading to the failure of

visible light imaging and difficulties in small target detection, as well as the high resource consumption hindering model deployment, an improved algorithm YOLOv8-RGBT-MT is proposed.

First, the FEM is employed, integrating the visible RGB and thermal infrared images and mitigating issues of missed and false detections caused by the failure of visible light and infrared imaging, particularly under adverse weather conditions such as dust, fog, rain and snow. Second, the C2f-MTB, an enhanced version of the C2f module, is introduced into the backbone network to expand the receptive field of the model and improve the feature extraction capabilities of the network in complex backgrounds. Finally, the MFF mechanism is introduced to effectively integrate complementary information from visible and infrared images through an adaptive weighting strategy, enhancing both detection accuracy and robustness for small targets. The network architecture of the enhanced algorithm is illustrated in Fig. 1.

2.2 FEM

The FEM employs fast Fourier transform (FFT) to decompose input images into their constituent frequency components. The real and imaginary parts of the frequency spectrum corresponding to low-frequency and high-frequency information respectively are processed independently through a basic convolutional block CBS (Conv-BN-SiLU). This design explicitly mitigates low-frequency blurring induced by adverse weather conditions by enhancing discriminative frequency components, while selectively suppressing detrimental low-frequency components and enhancing the contrast of informative low-frequency signals and high-frequency details, thereby improving feature extractability while preserving critical edge and texture details. The refined frequency representations are subsequently reconstructed into the spatial domain via inverse FFT (IFFT) and further optimized through an additional CBS module. Finally, the enhanced features are concatenated with the original input through a residual connection to enable complementary feature fusion. This module enables effective visible image enhancement and noise suppression, particularly under challenging atmospheric conditions including dust, fog, rain, and snow. The network structure of the FEM is shown in Fig. 2.

2.3 C2f-MTB

Mamba, an emerging architecture based on structured state space sequence models (SSMs), incorporates advantageous characteristics of both recurrent neural networks (RNNs) and convolu-

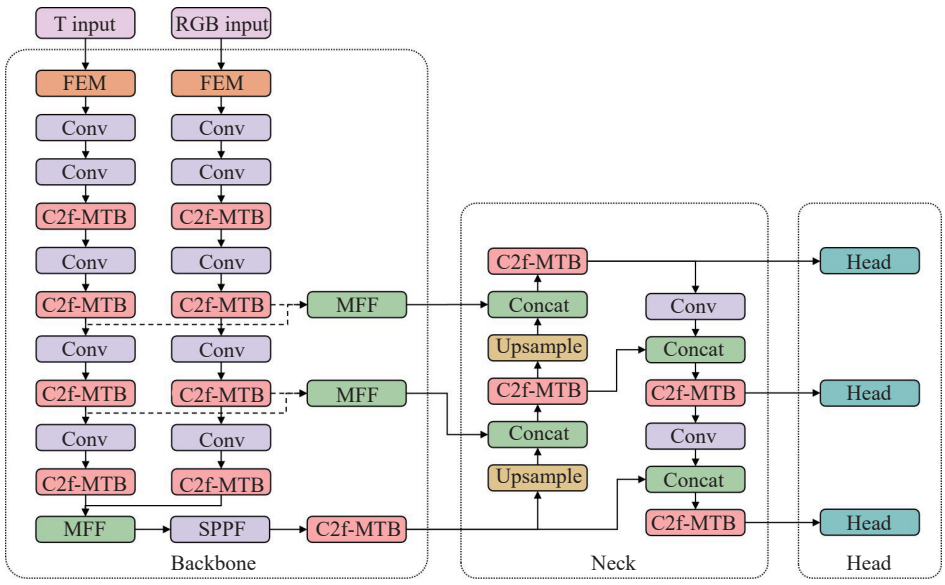

Fig. 1 Network structure of the improved algorithm

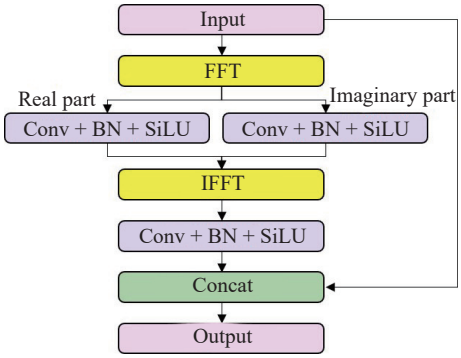

Fig. 2 FEM structure

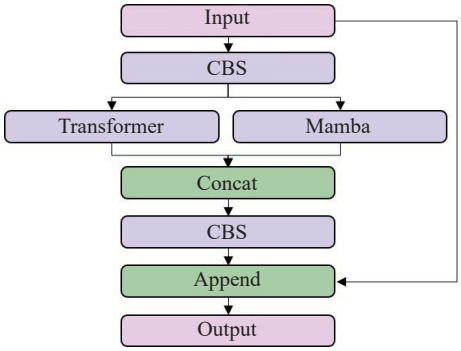

Fig. 3 MT-Bottleneck structure

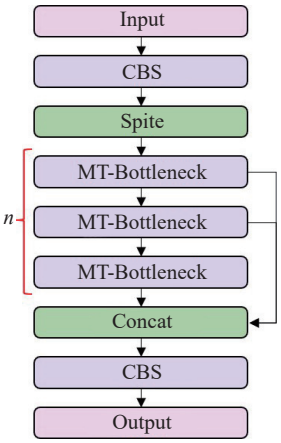

Fig. 4 C2f-MTB structure

tional neural networks (CNNs). It effectively captures local features while achieving linear or near-linear computational scaling with sequence length through recursive or convolutional operations, thereby significantly reducing computational complexity [23]. In contrast, the multi-head self-attention (MHSA) mechanism in transformers excels at modeling global dependencies. Owing to its powerful capacity for capturing global contextual information, the transformer architecture has attracted considerable attention in computer vision tasks. Nevertheless, its quadratic computational complexity relative to sequence length results in substantial computational overhead [24].

To address this limitation while maintaining effective feature extraction, this paper introduces a Mamba-Transformer hybrid Bottleneck (MT-Bottleneck). MT-Bottleneck partitions the input feature map by splitting the channel dimension and processing different channel subsets through dedicated Mamba and Transformer branches [25]. This design enables simultaneous exploitation of both local and global representations, effectively combining the complementary strengths of Mamba and Transformer to enhance the overall modeling capability. The network structure of the MT-Bottleneck is shown in Fig. 3 [26].

The original Bottleneck module in the C2f structure is replaced with the proposed MT-Bottleneck, forming an enhanced C2f-MTB architecture as shown in Fig. 4. By processing feature maps through parallel Mamba and Transformer branches within the MT-Bottleneck, the network achieves simultaneous extraction of both local and global features. This architectural modification enhances the model's feature representation capacity while effectively expanding the receptive field beyond the original configuration.

2.4 MFF Module

The visible RGB image offers rich color, texture, and fine-grained details, whereas the thermal infrared image relies on thermal radiation characteristics, making it invariant to illumination variations and robust under adverse weather conditions. The MFF mechanism effectively integrates these complementary modalities by employing an multilayer perceptron (MLP) to autonomously learn scenario-specific importance weights for each modality. The MLP is tasked with measuring the importance of features from two images, mitigating overfitting through a compact intermediate layer. Its output employs a Sigmoid function to compute weights for the two chan-

nels. The training process is conducted in two stages: first, the MLP is frozen while the detection network is trained, followed by unfreezing the MLP and fine-tuning it with a reduced learning rate. This strategy prevents the weights from biasing toward a single modality, thereby ensuring balanced contributions from both. These weights are normalized via a softmax function to generate attention coefficients. Using this adaptive strategy, the fusion mechanism enhances discriminative feature representation for improved target detection. Finally, the refined features are combined with the original input through a residual connection, enabling complementary fusion of high-level semantic information and low-level spatial features, thereby boosting both detection accuracy and overall system robustness. The network structure of the MFF module is shown in Fig. 5.

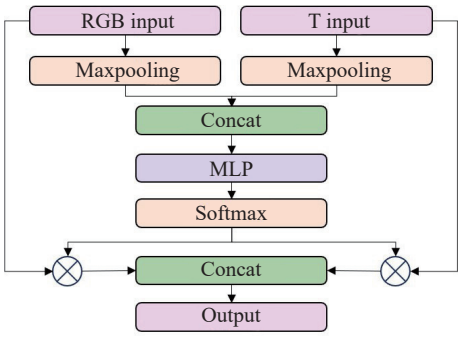

Fig. 5 MFF structure

3 Experimental Results and Analysis

3.1 Dataset and Environment Configuration

To evaluate the effectiveness of the proposed model, this study employs two distinct datasets: the high altitude UAV multi-modal tracking (HiAl) Dataset [27] and the severe weather visible-infrared (SWVI) dataset [28]. The HiAl dataset delivers 12 000 high-resolution frames gathered above construction sites, urban streets, lakesides, plazas and further locales, spanning ten-plus scenario categories and multiple illumination conditions of sunny daylight, night-time and dense fog. From this dataset, the excavator category is selected as the primary detection target. The SWVI dataset comprises 9 500 keyframes extracted from 80 surveillance video sequences. This dataset is particularly suitable for multi-object detection and tracking in complex environments. In this work, the person category under adverse weather conditions is adopted as the primary recognition target from SWVI.

For unbiased evaluation, the two datasets are trained completely independently, and no cross-dataset weight sharing or mixed sampling is performed. For each dataset, the official split is adopted.

All experiments were conducted within the Python environment. Tab. 1 shows the main experimental equipment environment configuration.

Tab. 1 Experimental environment configuration

Environment configuration	Version
Operating system	Ubuntu 18.04.6 LTS
Central processing unit (CPU)	Intel(R) Xeon(R) Gold 6240 CPU @ 2.60 GHz
GPU	VIDIA A100-PCIE-40GB
Python	3.11.5
CUDA	11.8

3.2 Evaluation Metrics

In this study, we adopt precision (${P}$), recall (${R}$), mean average precision at 50% IoU threshold (mAP50) and over 50%–95% IoU thresholds (mAP50–95) as the primary metrics for evaluating model performance. The mathematical definitions of these metrics are given as follows:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{mAP} = \frac{\sum_{i=1}^{n} \mathrm{AP}_i}{n} \tag{3}$$

where TP denotes the number of correctly

detected objects, FP indicates the number of erroneous detections, and FN refers to the count of undetected ground truth objects. Here, n represents the total number of target object classes, and AP characterizes the detection accuracy for each individual category.

3.3 Ablation Experiments

To evaluate the individual contributions of the proposed modules in multi-modal image detection, four ablation experimental configurations were established based on the YOLOv8s architecture. As summarized in Tab. 2 and Tab. 3, baseline model represents the YOLOv8-RGB model. Subsequent configurations sequentially integrate the FEM and the C2f-MTB module respectively to systematically assess their impacts.

Tab. 2 Ablation experiment results in HiAI (excavator category)

Methods	FEM	C2f-MTB	P(%)	R(%)	mAP50(%)	mAP50–95(%)
Ours	√	√	**91.4**	**71.5**	**86.5**	**57.2**
Baseline+ C2f-MTB		√	91.1	73.1	85.4	54.2
Baseline+ FEM	√		90.6	62.3	74.5	50.7
Baseline			85.9	58.8	67.8	37.9

Tab. 3 Ablation experiment results in SWVI (person category)

Methods	FEM	C2f-MTB	P(%)	R(%)	mAP50(%)	mAP50–95(%)
Ours	√	√	**88.1**	**81.9**	**88.4**	**52.3**
Baseline+ C2f-MTB		√	86.8	81.5	88.1	51.9
Baseline+ FEM	√		86	79.2	87.3	51.3
Baseline			85.2	78.9	86.4	50.7

Results from Tab. 2 and Tab. 3 quantitatively indicate that both the FEM and C2f-MTB modules enhance detection performance, with the latter yielding more pronounced improvements. Integrating the C2f-MTB module into the baseline brings absolute mAP50 gains of 17.6% on the HiAl dataset (excavator detection) and 1.7% on the SWVI dataset (person detection under adverse weather), with corresponding mAP50–95 improvements of 16.3 % and 1.2 %, respectively. Similarly, incorporating the FEM module leads to mAP50 increases of 6.7% (HiAl) and 0.9% (SWVI), with mAP50–95 rising by 12.8% and 0.6%. Furthermore, the combined integration of both modules yields additional performance gains, attributable to their complementary nature: the FEM enhances visible image quality under adverse weather conditions such as dust, fog, rain, and snow, while the C2f-MTB extends the model's receptive field and improves feature representation in complex scenes.

Overall, the ablation experiments validate that the proposed modules significantly boost detection accuracy for construction site objects, even under challenging environmental conditions.

3.4 Comparison of Different Approaches

While the proposed YOLOv8-RGBT-MT architecture integrates novel FEM and C2f-MTB modules to leverage complementary advantages in local feature extraction and global contextual modeling, thereby enhancing feature representation under adverse weather conditions. Its increased structural complexity may introduce potential trade-offs in model compactness. To comprehensively assess these trade-offs, we compare the improved model against the YOLOv8-RGBT baseline across multiple efficiency metrics: parameter count (Params), computational complexity (GFLOPs), model size, and inference time, as summarized in Tab. 4.

Tab. 4 Comparison between different algorithms

Index	Baseline	Ours
Params(M)	4.52	4.77
GFlops	12.18	12.23
Model Size(MB)	9.01	9.87
Inference time(ms)	406.2	82.3

Experimental results reveal a notable reduction in inference time, attributable to the efficient GPU-oriented computation of the Mamba component within the C2f-MTB module, which replaces the original C2f structure. Although, the integration of multiple new modules leads to a corresponding increase in GFLOPs, model size

and parameter count, this marginal expansion remains justifiable given the substantial gains in detection performance and operational speed.

As shown in Tab. 2 and Tab. 3, the experiments validate that the proposed approach boost detection accuracy for construction site objects, even under challenging environmental conditions. The visualization results of excavator detection are presented in Fig. 6 (a) and the visualization results of person detection are presented in Fig. 6(b), where (1) depicts the true annotations, while (2) and (3) present the detection results obtained by our improved algorithm and the baseline YOLOv8-RGB model, respectively.

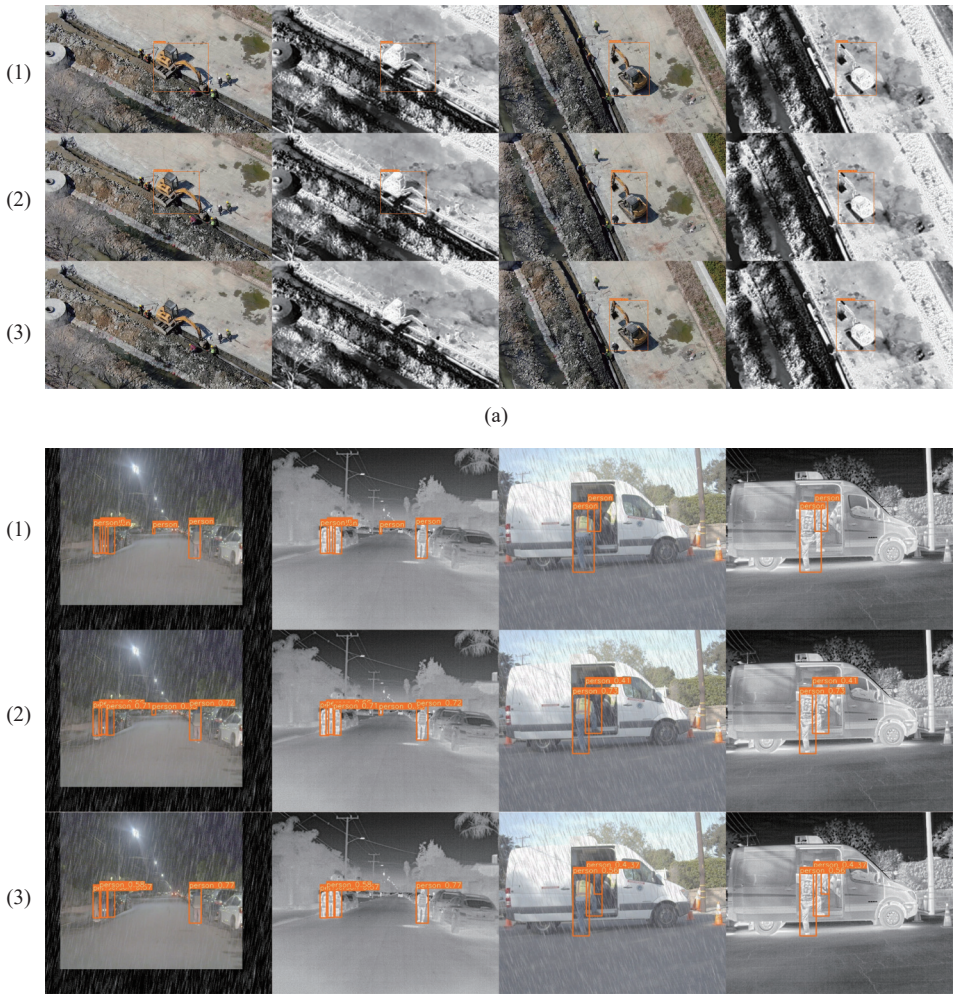

Fig. 6 Comparisons of different algorithms: (a) excavator detection; (b) person detection

4 Conclusion

This study presents the YOLOv8-RGBT-MT model, an efficient and computationally optimized algorithm for construction site object detection. The proposed method incorporates the FEM to improve the quality of visible and infrared images under adverse atmospheric conditions such as dust, fog, rain and snow, enabling more reliable feature extraction across varying weather scenarios. Furthermore, it introduces an enhanced C2f-MTB module that expands the receptive field of the model, which significantly strengthens its capability to extract discriminative features in complex backgrounds. The MFF mechanism is adopted to effectively integrate complementary information from visible and infrared images through an adaptive weighting

strategy, enhancing both detection accuracy and robustness for small targets. Experimental evaluation on the HiAI dataset demonstrates that the improved model achieves a mAP50 of 86.5%, corresponding to an 18.7% improvement over the original YOLOv8-RGB model, while reducing inference time by 79.7%. These results confirm that the proposed algorithm achieves substantial performance gains in construction site object detection and outperforms the YOLOv8-RGB model.

References:

[1] H. Liu, J. Song, and G. Wang,"A scientometric review of smart construction site in construction engineering and management: Analysis and visualization," *Sustainability*, vol. 13, no. 16, pp. 8860-8860, 2021.

[2] D. Qi, W. Yu, and Q. Chen, "MC-YOLO: A context aware down sampling algorithm for construction site object detection," in *2025 4th Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, pp. 1-6, 2025.

[3] S. Zhao, K. Kang, C. Xu, X. Guo, and R. Y. Zhong, "Digital twin enabled construction site monitoring (CSM) method with edge-cloud collaboration," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*, pp. 3017-3022, 2024.

[4] M. Mohammadi, M. Salimi, M. Loni, and S. Sinaei, "Enhancing object detection for autonomous machines in private construction sites through federated learning," in *2024 13th International Conference on Computer Technologies and Development (TechDev)*, pp. 39-43, 2024.

[5] Y. Wang, "Research on a safety helmet detection method based on smart construction site," in *2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, pp. 341-343, 2021.

[6] H. Zeng, X. Wu, Y. Lyu, and C. Yu, "Research on safety hazard identification method for construction sites based on YOLOv9 algorithm," in *2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML)*, pp. 451-455, 2024.

[7] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "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, 2023.

[8] J. E. Espinosa-Oviedo, S. A. Velastin-Carroza, and J. W. Branch-Bedoya,"EspiNet V2: A region based deep learning model for detecting motorcycles in urban scenarios," *Dyna (Medellin, Colombia)*, vol. 86, no. 211, pp. 317-326, 2019.

[9] S. Ren, K. He, R. Girshick, and J. Sun,"Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.

[10] X. Huang, Y. Jiang, G. Tian, and C. Duan, "Research on small object detection algorithm in complex construction environment based on deformable convolutional operator," in *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)*, pp. 1-6, 2024.

[11] X. Guo, H. Liu, Y. Zhao, P. Dong, Y. Wang, X. Li, and Q. Zhuang, "YOLO-MSD: An accurate detection method for small target on construction sites," in *2025 37th Chinese Control and Decision Conference (CCDC)*, pp. 1-7, 2025.

[12] R. Ge and S. Cong, "Target detection algorithm for construction sites with high background noise based on improved YOLO," in *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)*, pp. 891-895, 2025.

[13] Y. Sun, L. Zheng, Z. Lu, Q. Lian, and L. Zhou, "A bolt rust detection system for smart construction sites," in *2023 8th International Conference on Image, Vision and Computing (ICIVC)*, pp. 403-408, 2023.

[14] H. Wang, Q. He, J. Peng, H. Yang, M. Chi, and Y. Wang, "Mamba-YOLO-World: Marrying YOLO-world with Mamba for open-vocabulary detection," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2025.

[15] S. Wu, X. Lu, C. Guo, and H. Guo,"MV-YOLO: An efficient small object detection framework based on Mamba," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1-14, 2025.

[16] J. Wang, W. Zhao, C. Liu, H. Yang, and W. Xu, "Real-time object detection based on Mamba and

YOLOv8," in *2024 4th International Conference on Industrial Automation, Robotics and Control Engineering (IARCE)*, pp. 255-260, 2024.

[17] J. An, P. Tao, M. Dwisnanto Putro, and B. W. Kim,"Mamba*YOLO: Lightweight and accurate object detection via regional attention with gated enhancement," *IEEE Access*, vol. 13, pp. 175495-175518, 2025.

[18] Y. Dai, W. Liu, H. Wang, W. Xie, and K. Long,"YOLO-Former: Marrying YOLO and Transformer for foreign object detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-14, 2022.

[19] Z. Chen, K. Yang, Y. Wu, H. Yang, and X. Tang,"HCLT-YOLO: A hybrid CNN and lightweight transformer architecture for object detection in complex traffic scenes," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 3, pp. 3681-3694, 2025.

[20] L. Cheng,"A highly robust helmet detection algorithm based on YOLO v8 and Transformer," *IEEE Access*, vol. 12, pp. 130693-130705, 2024.

[21] W. Zhou, C. Cai, C. Li, H. Xu, and H. Shi,"AD-YOLO: A real-time YOLO network with Swin Transformer and attention mechanism for airport scene detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1-12, 2024.

[22] L. Qin, W. Mei, H. Cui, H. Bian, and X. Wang,"Improved YOLOv8s detection algorithm for remote sensing images," *Journal of Beijing Institute of Technology*, vol. 34, no. 3, pp. 278-289, 2025.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv*, vol. abs/2010.11929, 2020.

[24] Y. Fang, B. Liao, X. Wang, J. Fang, J. Qi, R. Wu, J. Niu, and W. Liu, "You only look at one sequence: Rethinking transformer in vision through object detection," in *Advances in Neural Information Processing Systems*, ser. NIPS '21, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, pp. 26183-26197, 2021.

[25] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *ArXiv*, abs/2312.00752, 2023.

[26] Z. Wang, C. Li, H. Xu, X. Zhu, and H. Li, "Mamba YOLO: A simple baseline for object detection with state space model," *arXiv e-prints*, arXiv: 2406.05835, 2024.

[27] Y. Xiao, D. Cao, C. Li, B. Jiang, and J. Tang,"A benchmark dataset for high-altitude UAV multimodal tracking," *Journal of Image and Graphics*, vol. 30, no. 2, pp. 361-374, 2025.

[28] H. Li, Q. Hu, B. Zhou, Y. Yao, J. Lin, K. Yang, and P. Chen, "CFMW: Cross-modality fusion mamba for robust object detection under adverse weather," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 12, pp. 12066-12081, 2025.

Yan Li received the Bachelor's degree in Engineering Management from Hunan University in 2008, and the Master's degree in Management from Renmin University of China in 2015. He has been pursuing his Ph.D. in Control Engineering from 2018 at Beijing Institute of Technology. His research interests focus on prefabricated buildings, smart buildings, as well as intelligent information processing and control.

Cunxin Sun is currently pursuing the B.Eng. degree in Automation at Beijing Institute of Technology. His current research interests include computer vision and robust model predictive control.

Baihai Zhang received his Ph.D. from Harbin Institute of Technology in 1994. Since July 2002, he is a professor at Beijing Institute of Technology. His current research interests focus on wireless sensor networks and industrial Internet, theory and applications of systems engineering, as well as modeling, simulation and optimization of complex mechatronic systems.