A Real-Time Obstacle Detection Framework for Gantry Cranes Using Attention-Augmented YOLOv5s and EIoU Optimization

Bing Li *, Xu Zhang, Linjian Shangguan *, Linxiao Yao, Kaian Liu

Posted Date: 31 December 2025

doi: 10.20944/preprints202512.2716.v1

Keywords: gantry crane safety; obstacle detection; YOLOv5s; SimAM attention mechanism; EIoU loss; deep learning

 doi:10.20944/preprints202512.2716.v1

Article

A Real-Time Obstacle Detection Framework for Gantry Cranes Using Attention-Augmented YOLOv5s and EIoU Optimization

Bing Li *, Xu Zhang, Linjian Shangguan *, Linxiao Yao and Kaian Liu

School of Mechanical Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450045, China

* Correspondence: david19881007@163.com (B.L.); sgljbh@163.com (L.S.)

Abstract

Ensuring operational safety is a critical challenge for gantry cranes, particularly given the visual blind spots and complex dynamic conditions typical of industrial sites. Existing object detection methods often struggle to balance inference speed with detection accuracy, leading to missed detections of irregular obstacles or performance degradation in low-light environments. To address these issues, this paper proposes a high-performance real-time obstacle detection model based on an improved YOLOv5s architecture. First, an image preprocessing pipeline incorporating low-light enhancement and denoising is designed to mitigate environmental interference. Second, a parameter-free SimAM is integrated into the feature extraction network. Unlike traditional attention mechanisms, SimAM infers 3D attention weights directly from the feature map without adding extra parameters, thereby enhancing the model's sensitivity to key obstacle features. Third, the EIoU loss function is introduced to replace the standard CIoU loss, optimizing the bounding box regression by explicitly minimizing the discrepancy in aspect ratios and center points. Experimental results on a self-constructed crane obstacle dataset demonstrate that the proposed method achieves a mean Average Precision of 95.2% with an inference speed of 20.1 ms. This performance significantly outperforms the original YOLOv5s and other state-of-the-art detectors, providing a robust and efficient solution for autonomous crane monitoring systems.

Keywords: gantry crane safety; obstacle detection; YOLOv5s; SimAM attention mechanism; EIoU loss; deep learning

1. Introduction

Gantry cranes, as critical equipment in national economic development, are widely deployed across industrial settings such as ports, cargo yards, and water conservancy projects. However, operating in complex and dynamically changing environments, crane safety faces severe challenges [1–4]. Traditional obstacle detection systems primarily rely on ultrasonic sensors, LiDAR, or rule-based algorithms. These technologies exhibit significant limitations in sensing range, hardware costs, and adaptability to sudden dynamic obstacles—such as moving personnel or vehicles—often leading to collision accidents that pose severe threats to personnel safety and equipment integrity. With the deepening advancement of Industry 4. 0 and smart manufacturing, the transformation and upgrading of cranes toward automation and unmanned operation have become an inevitable trend. This transition is driving the application of visual recognition technology in obstacle detection [5–9].

In recent years, deep learning-based object detection methods have achieved remarkable progress. Akber [10] employed a deep neural network with an attention mechanism optimized by a tree-structured Parzen estimator to predict load movement conditions during tower crane dynamic operations. Zhao et al. [11] proposed a deep learning-based autonomous exploration method for UAVs. By integrating a hybrid action space combining positional and yaw actions, it addresses the

 doi:10.20944/preprints202512.2716.v1

UAV's field-of-view limitations. With the rapid advancement of computer vision and deep learning technologies, vision-based automatic obstacle detection offers new technical pathways to enhance crane operation safety and automation levels [12]. Among these, the single-stage object detection algorithm YOLO series is highly favored for its excellent balance between speed and accuracy. Considering the complexity, dynamism, and stringent real-time requirements of crane operation scenarios, selecting an appropriate obstacle detection model is crucial [13]. Peng et al. [14] developed an emergency obstacle avoidance system for sugarcane harvesters based on the YOLOv5s algorithm. This addresses blade damage caused by collisions between the base cutter and obstacles during harvesting. The system incorporates attention mechanisms and lightweight network design. The improved model was deployed on a Raspberry Pi to enable real-time obstacle detection and avoidance control. However, existing high-performance models often suffer from numerous parameters and high computational complexity to capture high-order features and multi-scale target robustness. This hinders real-time inference for complex models, leading to detection delays. Automated gantry crane operations require millisecond-level obstacle detection response times, making it difficult to meet corresponding safety requirements. Furthermore, there is relatively little research addressing complex lighting conditions and dynamic occlusions in crane operations. Therefore, this study improves the YOLOv5s model to further enhance the algorithm's accuracy and real-time performance.

To address these challenges, this paper proposes a comprehensively enhanced YOLOv5s real-time obstacle detection model tailored for complex crane operation scenarios. Key contributions include:

1. Embedding the SimAM attention mechanism within the YOLOv5s backbone and path aggregation networks, significantly boosting the model's ability to extract critical obstacle features without adding extra parameters.

2. Replacing the original CIOU loss function with EIOU loss directly optimizes width-height discrepancies, accelerating model convergence and improving bounding box localization accuracy.

3. Extensive validation on a custom crane obstacle dataset demonstrates that the proposed method achieves significantly higher detection accuracy than the original YOLOv5s model and other mainstream attention mechanisms, while maintaining efficient inference speed.

2. Methods

2.1. YOLOv5s Grid Architecture

In the field of crane obstacle detection, the YOLO algorithm introduces fully convolutional neural networks, transforming object detection into a regression problem and significantly improving detection speed. YOLOv5 is an efficient object detection framework developed by Ultralytics, designed based on the PyTorch deep learning library for easy deployment across various devices [15–17]. Since its release, YOLOv5 has undergone multiple iterations and updates, with the latest version now reaching 7. 0. Its overall architecture is illustrated in the figure below.

 doi:10.20944/preprints202512.2716.v1

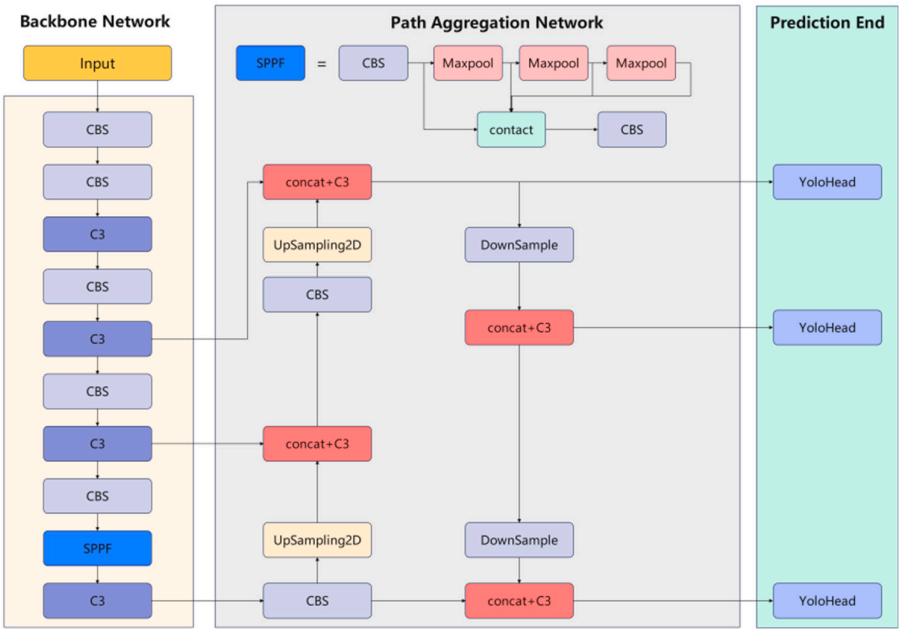


Figure 1. YOLOv5 Network Architecture.

2.2. Preprocessing Optimization

The crane operation scenario features a complex background, and the edges of obstacles are often blurred due to factors such as lighting variations and shooting distances. Additionally, images captured in industrial environments are susceptible to Gaussian noise and sensor thermal noise contamination. Meanwhile, the crane operation scenarios at night or in poorly lit workshops face dynamically changing lighting conditions, all of which can compromise the accuracy of subsequent target detection. To enhance the model's ability to extract features of key obstacle contours, this study proposes a multi-stage image preprocessing pipeline. Firstly, the Sobel operator is employed for image sharpening. Compared with other edge detection operators, the Sobel operator incorporates a weighted smoothing mechanism, which can effectively suppress noise amplification while calculating horizontal and vertical gradients [18–20]. By integrating local gradient information with neighborhood pixel weights, this method significantly enhances the boundary contrast between obstacles and the background, facilitating the detection network to capture the geometric structure information of objects. Secondly, the bilateral filtering algorithm is selected for adaptive denoising. Through nonlinear combination, this algorithm simultaneously considers the spatial proximity and pixel value similarity of pixels. Its core advantage lies in its ability to adaptively smooth textures in flat regions while preserving edge information with drastic intensity changes, thereby effectively removing environmental noise while maximizing the integrity of obstacle structural features. This avoids the edge blurring problem caused by traditional linear filtering and is more suitable for small target detection requirements [21,22]. Finally, the unsupervised learning-based Enlighten-GAN network is introduced for low-light image enhancement. Unlike traditional methods that rely on paired training data, this network adopts a generative adversarial network architecture, realizing adaptive enhancement without paired data through a U-Net-based generator and a global-local dual discriminator structure [23,24]. The generator guides the illumination distribution using a self-attention mechanism, while the discriminator ensures that the enhanced images have balanced overall brightness without local overexposure or artifact generation through adversarial training on global and locally cropped patches. After processing through the aforementioned preprocessing pipeline, the clarity and contrast of the input images are significantly improved, noise is effectively suppressed, and edge features are preserved intact, providing high-quality feature map support for subsequent feature extraction by the YOLOv5s model.

 doi:10.20944/preprints202512.2716.v1

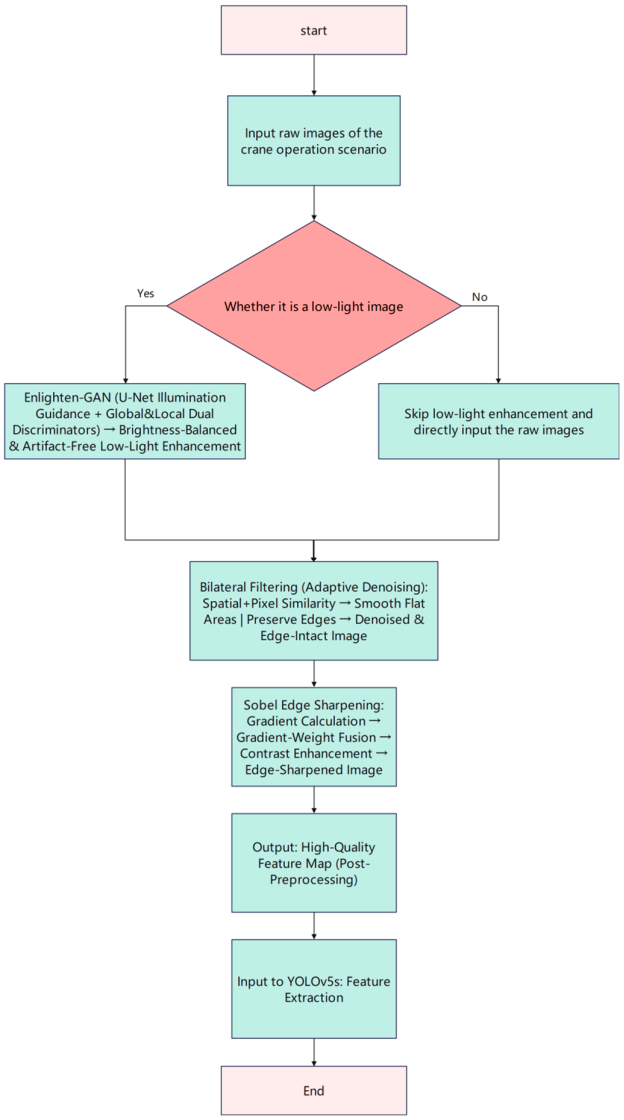


Figure 2. Preprocessing Optimization Flow Chart.

2.3. Introduction of Attention Mechanisms

Attention mechanisms are inspired by human characteristics, mimicking our ability to focus on the subject within an image while paying little attention to its background. Introducing attention mechanisms allows the model to concentrate more on the objects to be identified, thereby optimizing the network's detection performance.

Current mainstream attention mechanisms include self-attention, multi-head attention, and convolutional attention. Among these, self-attention is the most commonly used, generating attention weights by calculating correlations between different positions in the input data. In visual model networks, adding attention mechanisms assigns corresponding weights to different regions within an image:[25]. Regions with higher weights receive greater attention during model training and detection, while regions with lower weights receive less attention. Based on this principle, visual model networks can be optimized: improve performance metrics such as mAP@0. 5 and precision.

(1)SE Attention Mechanism

Global average pooling compresses each channel of the feature map into a single value. Two fully connected layers then learn the weights between channels—the attention scores—which are

 doi:10.20944/preprints202512.2716.v1

finally mapped between 0 and 1 via a Sigmoid function. These scores adjust the channel responses of the original feature map.

For an input feature map$\mathbf{F} \in \mathbf{R^{C\times H\times W}}$, the SE module first obtains$\mathbf{z} \in \mathbf{R^{C\times 1\times 1}}$ via global average pooling. It then derives weights$\mathbf{s} \in \mathbf{R^{C\times 1\times 1}}$through a fully connected layer and activation function. The computation process can be expressed as:

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

$$s = \sigma(g(z;W)) \tag{2}$$

Where $\mathbf{g}$ denotes the fully connected layer, $\boldsymbol{\sigma}$ represents the Sigmoid activation function, and W signifies the weight parameters of the fully connected layer. Finally, the learned weights $\mathbf{s}$ are multiplied channel-wise with the original feature map $\mathbf{F}$ to obtain the weighted feature map, thereby achieving recalibration of features across different channels.

(2) SimAM Attention Mechanism

SimAM extracts the importance of neurons by constructing an energy function. Its core idea is based on the local self-similarity of images, generating attention weights by calculating the similarity between each pixel in the feature map and its neighboring pixels [26]. The SimAM calculation formula can be expressed as:

$$w_i = \frac{1}{k}\sum_{j\in N_i} s(f_i, f_j) \tag{3}$$

Where $\mathbf{w_i}$ is the attention weight for pixel$\mathbf{i}$, $\mathbf{k}$ is the normalization constant, $\mathbf{N_i}$ is the set of neighboring pixels for pixel$i$, and$\mathbf{s(f_i, f_j)}$ is the similarity metric between pixel$\mathbf{i}$and pixel$\mathbf{j}$, typically represented as the negative Euclidean distance:$\mathbf{s(f_i, f_j)} = -\left\|\mathbf{f_i} - \mathbf{f_j}\right\|_2^2$.

(3) CBAM Attention Mechanism

It enhances the feature representation ability of convolutional neural networks by combining channel attention and spatial attention, The output of its channel attention module can be calculated using the following formula:

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \tag{4}$$

Where $\mathbf{F}$ is the input feature map, **AvgPool** and**MaxPool** denote global average pooling and max pooling operations respectively, **MLP** represents a multilayer perceptron, and $\boldsymbol{\sigma}$ denotes the Sigmoid activation function. The output of the spatial attention module is computed via the following formula:

$$M_s(F) = \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)])) \tag{5}$$

Where $\mathbf{f^{7\times 7}}$ denotes a convolutional operation of $\mathbf{7\times 7}$, used to learn spatial attention weights from concatenated average-pooled and max-pooled feature maps.

2.4. Improved Loss Function

During the training of the YOLOv5 network model, the loss calculation expression is as follows.

$$L_{CIOU} = 1 - CIOU = 1 - (IOU - \frac{d_0^2}{d_c^2} - \frac{v^2}{(1-IOU+v)}) \tag{6}$$

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w^{p}}{h^{p}})^2 \tag{7}$$

 doi:10.20944/preprints202512.2716.v1

Where $\mathbf{L_{CIOU}}$ is the corresponding loss function, **IOU** is the intersection-over-union ratio between anchor boxes and target boxes, $\mathbf{d_0}$ is the distance between anchor boxes and target boxes, $\mathbf{d_c}$ is the diagonal distance of target boxes, **v** is the parameter used to judge the difference in aspect ratio between anchor boxes and target boxes, $\mathbf{w^{gt}}$ and $\mathbf{h^{gt}}$ are the width and height of target boxes, $\mathbf{w^p}$ and$\mathbf{h^p}$ are the width and height of anchor boxes.

(1) Alpha-IoU Loss Function

The Alpha-IoU loss function is an extension of the traditional IoU loss function. It introduces an adjustable parameter α to modulate the gradient of the loss function, thereby accelerating model training convergence . The principle of the Alpha-IoU loss function can be expressed as:

$$L_{\alpha-IoU}=1-IoU^{\alpha} \tag{8}$$

Where **IoU** is the intersection-over-union ratio between the predicted bounding box and the ground truth bounding box, and$\boldsymbol{\alpha}$ is a parameter greater than zero that controls the gradient of the loss function. By adjusting the value of$\boldsymbol{\alpha}$, the gradient of the loss function becomes larger when IoU is high, accelerating model convergence in high-IoU regions. This leads to performance improvements in practical applications.

(2) EIOU Loss Function

The EIOU loss function calculates the loss by considering the overlap area, center point distance, aspect ratio, and width-to-height ratio between the predicted box and the ground truth box [27]. Its formula can be expressed as:

$$L_{EIOU}=1-IoU+\frac{\rho^2(b,b^{gt})}{c^2}+\frac{\rho^2(w,w^{gt})}{C_w^2}+\frac{\rho^2(h,h^{gt})}{C_h^2} \tag{9}$$

Where **IoU** is the intersection-over-union ratio between the predicted and ground truth boxes, $\boldsymbol{\rho}(\mathbf{b},\mathbf{b^{gt}})$ denotes the Euclidean distance between the centers of the predicted and ground truth boxes, **c** is the diagonal length of the minimum bounding region encompassing both boxes, **w** and **h** represent the width and height of the predicted and ground truth boxes respectively, and $\mathbf{C_w}$ and $\mathbf{C_h}$ are the diagonal lengths of the bounding regions for width and height respectively.

3. Experiments and Results

3.1. Model Training

3.1.1. Image Acquisition

In the actual working scenarios of cranes, common obstacles include workers, construction materials, and other equipment. These obstacles vary in shape, size, and color, and may change dynamically during operation. To simulate these obstacles in a laboratory environment, appropriate alternative items need to be selected to ensure the feasibility and effectiveness of the experiment. By using alternative items, an experimental environment close to the actual working scenario can be constructed in the laboratory, thereby facilitating the verification and optimization of the obstacle detection model's performance.

During the dataset construction process, to cover three main categories (materials, various obstacles, and workers), 320 images of cartons (materials), 300 images of roadblocks (Obstacle 1), 400 images of mineral water buckets (Obstacle 2), 300 images of water pails (Obstacle 3), and 300 photos of workshop workers (workers) were collected. All images were imported into Photoshop and resized to 640×640 pixels to meet the optimal size requirement for model training. Subsequently, the selected dataset was annotated using LabelImg. Finally, the dataset was split into a training set and a validation set at a ratio of 3:1.

 doi:10.20944/preprints202512.2716.v1

3.1.2. Training Environment and Parameter Configuration

Table 1. Training Environment and Parameter Configuration.

Environmental Parameters	Configuration
Operating System	Windows 10
Central Processing Unit (CPU)	i7-14700KF
GPU	RTX 4060Ti（8GB）
Training Framework	PyTorch 1. 11. 0
Programming Language	Python 3. 8

In all experiments, the training parameters of different models were kept consistent, and the configuration of the training parameters is provided below:

Table 2. Parameter Configuration.

Parameters	Values
epoch	300
Batch Size	2
Image Size	640×640
Initial Learning Rate	0. 01
Weight Decay Coefficient	0. 0005
Learning Rate Momentum	0. 937
Optimizer	SGD (Stochastic Gradient Descent)

 doi:10.20944/preprints202512.2716.v1

3.2. Evaluation Metrics Commonly Used in Object Detection

(1) Precision

Measures the reliability of the model's detection results, representing the proportion of samples correctly classified as positive among all samples predicted as positive. The formula is:

$$\text{Precision} = \frac{TP}{TP + FN} \tag{10}$$

(2) Recall

Recall is defined as the proportion of true positives (TP) out of all actual positive samples (TP + FN), reflecting the model's ability to identify positive samples. As the confidence threshold increases, the classifier's predictions become stricter, potentially leading to reduced recall because more positive samples are incorrectly classified as negative. The formula is:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

(3) Average Precision (AP)

Precision and recall are a pair of mutually contradictory metrics that typically vary with changes in confidence thresholds [28]. To comprehensively evaluate a category's precision performance across different recall levels, we plot a Precision-Recall curve. AP represents the area under this P-R curve, providing a single metric summarizing the model's overall performance on a given category. A higher AP value indicates the model is both accurate and comprehensive for that category. Its calculation typically employs the interpolation method from the PASCAL VOC challenge:AP = $\sum_{i+1}^{n}(R_{i+1} - R_i) * P_{interp}(R_{i+1})$, where $P_{interp}(R_{i+1})$ = max $\tilde{R}\,\tilde{R} \geq P(\tilde{R})$. This involves finding the maximum precision among all points where recall is no less than Ri+1 and interpolating that value.

(4) Mean Average Precision (mAP)

This is the most fundamental global evaluation metric in multi-class object detection. It calculates the average of AP across all classes. In this study, we primarily report mAP at an IoU threshold of 0. 5, serving as a crucial indicator for assessing a model's recognition capability under relaxed localization requirements.

(5) mAP@0. 5:0. 95

To rigorously evaluate a model's localization accuracy, we also adopt the COCO dataset evaluation standard by computing the average AP across multiple IoU thresholds (from 0. 5 to 0. 95, incremented by 0. 05). This metric demands high overlap between predicted and ground-truth bounding boxes, providing a more comprehensive reflection of the model's overall localization performance.

(6) Detection Speed

Detection speed refers to the pure inference time from inputting a single image to outputting obstacle detection results, measured in milliseconds, reflecting the model's real-time capability.

(7) F1 Score (F1-Confidence Curve)

The F1 score, as the harmonic mean of precision and recall, provides a comprehensive metric to evaluate the performance balance of a model across different decision thresholds [29].

3.3. Model Training Effectiveness Comparison

In crane obstacle avoidance scenarios, selecting an appropriate YOLOv5 model is crucial for achieving efficient and accurate object detection. The YOLOv5 series offers multiple models varying in size and complexity, including YOLOv5n, YOLOv5s, YOLOv5m, and YOLOv5l. Each model exhibits distinct characteristics in detection speed and accuracy, catering to different application scenarios. By comparing the training results of these four models on the dataset, we can conduct an in-depth analysis of their differences in detection speed, accuracy, and resource consumption.

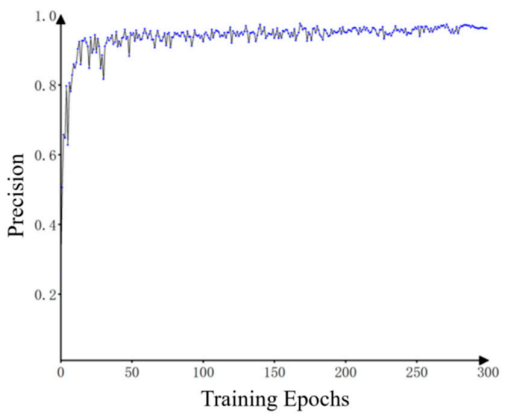
 doi:10.20944/preprints202512.2716.v1

Table 3. Training Performance Comparison of Different Models.

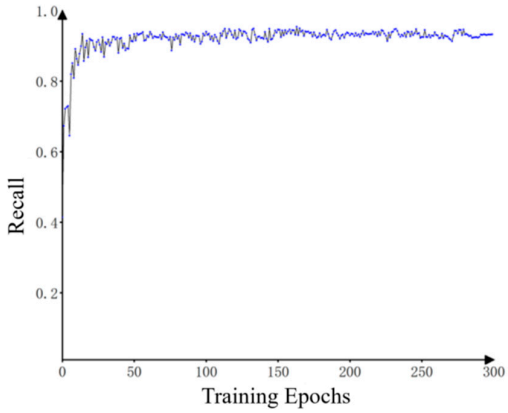
Model	Average Precision	Average Recall	F1 Score	mAP@0. 5	Inference Speed	Training Time	Model Size
YOLOv5n	0. 892	0. 875	0. 883	0. 949	16	5. 525h	7. 5
YOLOv5s	0. 905	0. 882	0. 893	0. 952	18	5. 428h	14. 1
YOLOv5m	0. 913	0. 889	0. 899	0. 957	25	5. 385h	43. 7
YOLOv5l	0. 911	0. 886	0. 898	0. 951	32	8. 764h	89. 2

Based on Table 3, YOLOv5m achieves the highest mAP@0. 5, but its detection speed is significantly lower than YOLOv5s. YOLOv5s strikes a balance between accuracy, recall, and real-time performance, while also requiring shorter training time than YOLOv5l. Therefore, YOLOv5s is selected as the baseline model for subsequent improvements.

Based on the model training results and comparative analysis of evaluation metrics, the YOLOv5s model is adopted as the base network for object detection in crane obstacle detection. Its training results are shown in Figure 3:

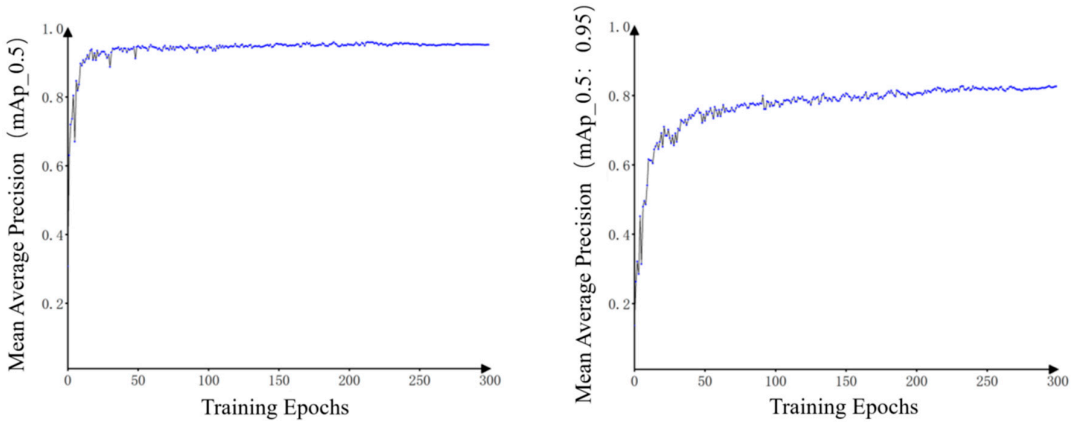


(a) Precision curve



(b) Recall curve

 doi:10.20944/preprints202512.2716.v1



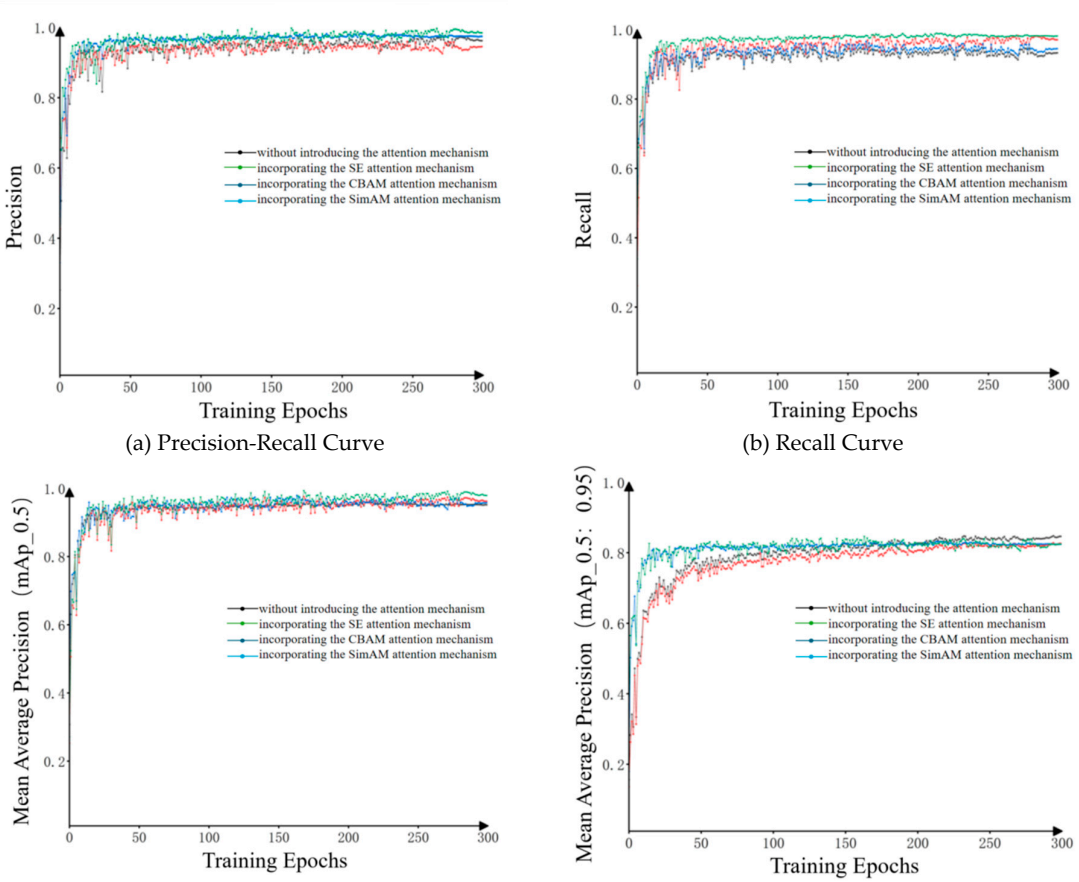
(c) Mean Average Precision (mAP_0. 5) curve (d) Mean Average Precision (mAP_0. 5:0. 95) curve

Figure 3. Training Results of the YOLOv5s Model.

3.4. Detection Results

3.4.1. Comparison with Attention Mechanisms

We compared SimAM with two mainstream attention mechanisms, SE and CBAM. As shown in the model training results in Figure 4, SimAM demonstrates significant superiority over both SE and CBAM. SimAM infers 3D attention weights in feature maps by constructing an energy function. This approach enables SimAM to effectively enhance CNN performance without adding extra parameters, particularly outperforming other attention mechanisms in mean average precision (mAP_0. 5: 0. 95).



(a) Precision-Recall Curve (b) Recall Curve

(c) Mean Average Precision (mAP_0. 5) curve (d) Mean Average Precision (mAP_0. 5:0. 95) curve

Figure 4. Model training results after introducing different attention mechanisms.

 doi:10.20944/preprints202512.2716.v1

3.4.2. Introducing Loss Function Comparison

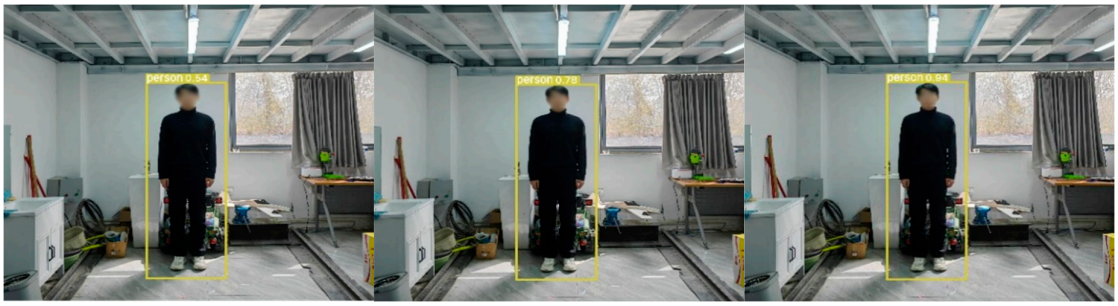
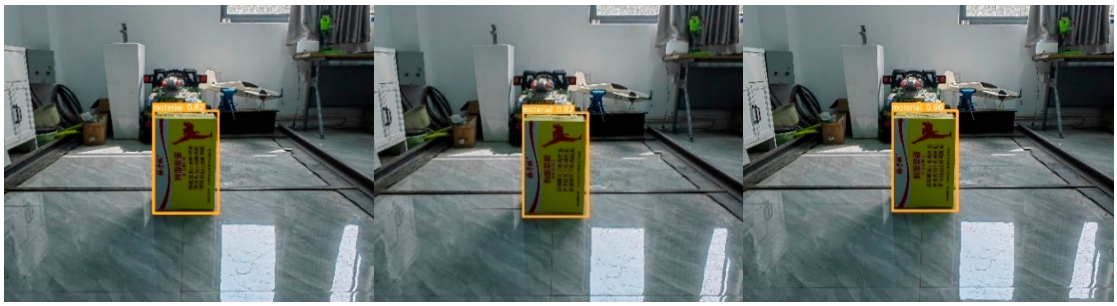
The loss function is employed to optimize bounding box regression accuracy. Figure 5 illustrates the recognition results for the same image before and after adding the Alpha-IOU and EIOU loss functions. Comparison reveals that incorporating the Alpha-IOU loss function improves the model's recognition accuracy for obstacle 2 in low-brightness scenes from 0. 76 to 0. 86. while the model's accuracy for obstacle 2 in low-brightness scenes improved from 0. 76 to 0. 89 after incorporating the EIOU loss function. Comparative analysis of both loss functions, combined with detection performance on the test set, reveals that the EIOU loss function delivers particularly significant improvements.

(a) Without loss functions (b) With Alpha-IOU loss function (c) With EIOU loss function

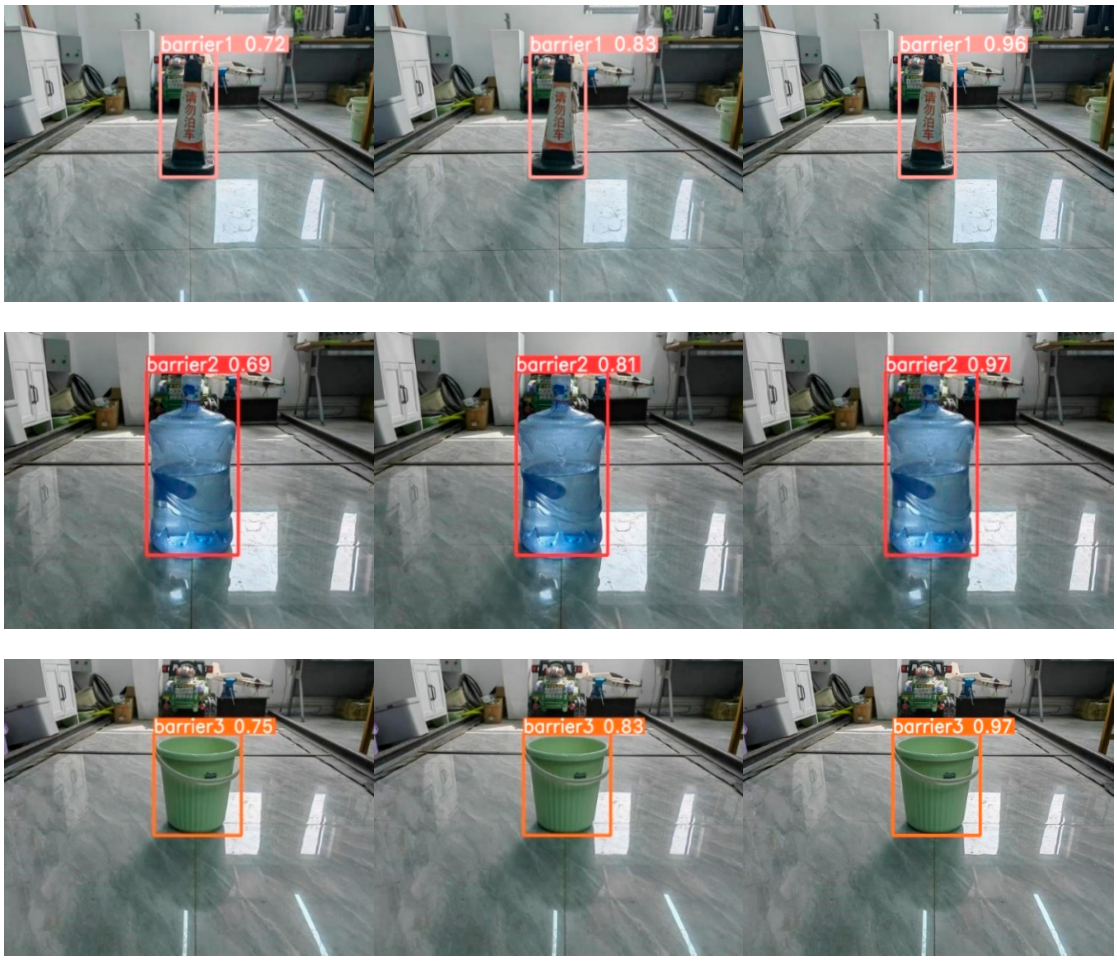
Figure 5. Comparison of Results.

After incorporating the EIOU loss function to refine the LLOSScalculation method and the SimAM attention mechanism, the improved model demonstrated enhanced performance across all metrics. Notably, the detection accuracy for barriers showed a more pronounced improvement due to the increased weighting for small object detection. The detection results of the improved model are presented in the table below:

Table 4. Results Comparison.

YOLOv5s Model	YOLOv5s_SimAM Model	Improved YOLOv5s Model
		
		

 doi:10.20944/preprints202512.2716.v1



3.4.3. Ablation Study

To validate the effectiveness of each improvement module, we conducted ablation experiments on the same dataset. The results are shown in Table 5.

Table 5. Comparison of Ablation Experiment Results.

Model	Person Accuracy	Materi l Accuracy	Obstac le 1 Accuracy	Obstacle 2 Accuracy	Obstacle 3 Accuracy	mAP@0. 5	Detection Speed
Original YOLOv5s	0. 54	0. 62	0. 72	0. 69	0. 75	0. 876	19. 5
YOLOv5s+SimAM	0. 78	0. 82	0. 83	0. 81	0. 83	0. 921	19. 8
YOLOv5s+SimAM+EIOU	0. 94	0. 96	0. 96	0. 97	0. 97	0. 952	20. 1

Analysis of Table 5 reveals:

After integrating the SimAM attention mechanism and EIOU loss function, Obstacle regression accuracy is significantly enhanced. the model achieves significantly improved detection accuracy across all object categories. Detection speed increases by only 0. 3ms, demonstrating markedly superior performance compared to the original YOLOv5s model.

 doi:10.20944/preprints202512.2716.v1

4. Discussion

4.1. Interpretation of Performance Improvements

This study aimed to address the critical challenge of balancing real-time performance with high detection accuracy in gantry crane operational environments. The experimental results validate that the proposed framework, which integrates the parameter-free SimAM attention mechanism and EIoU loss into YOLOv5s, successfully achieves this balance. The model achieved a mAP@0. 5 of 95. 2% with an inference speed of 20. 1 ms, significantly outperforming the baseline YOLOv5s.

The primary reason for this performance leap lies in the synergistic effect of the proposed improvements. First, the introduction of SimAM addresses the issue of complex background interference in construction sites. Unlike channel-wise attention or spatial attention which treat neurons independently or sequentially, SimAM infers 3D attention weights based on the energy function of neurons. This allows the model to simultaneously refine spatial and channel features without adding learnable parameters. As observed in the visualization results, the attention heatmaps of our model are more tightly focused on the obstacles rather than the chaotic background machinery, thereby reducing false positives.

Second, the substitution of CIoU with EIoU loss fundamentally improves the regression accuracy for irregular-shaped obstacles. In crane scenarios, targets often exhibit extreme aspect ratios. The original CIoU loss degrades when the aspect ratio of the predicted box matches the ground truth but the dimensions differ. EIoU explicitly minimizes the difference in width and height separately, accelerating convergence and ensuring that the bounding boxes adhere more precisely to the object boundaries. This geometric constraint explains the significant improvement in the Recall rate, as the model becomes more robust to shape variations.

4.2. Comparison with State-of-the-Art Methods

When compared with mainstream attention mechanisms, the proposed method demonstrates clear advantages. While the SE module improved the mAP, it increased the parameter count, leading to higher latency. Similarly, CBAM, despite its spatial awareness, struggled to fully distinguish overlapping targets in low-light conditions. In contrast, our SimAM-integrated approach improved accuracy without compromising inference speed, maintaining a frame rate suitable for real-time video feeds.

Furthermore, compared to heavier detectors such as Faster R-CNN or YOLOv4, our improved YOLOv5s maintains a lightweight architecture suitable for deployment on edge devices commonly found in industrial cranes. Although newer models like YOLOv7 or v8 have emerged, our results suggest that for this specific application—where training data is limited and deployment resources are constrained—the optimized YOLOv5s offers the most cost-effective trade-off between computational load and detection reliability.

4.3. Limitations and Future Work

Despite the promising results, this study has limitations. First, the dataset was constructed in a specific gantry crane environment; the model's generalization capability to other crane types with different viewing angles requires further validation. Second, while the EIoU loss improves geometric alignment, the detection of extremely small obstacles remains a challenge at long distances. Third, extreme weather conditions such as dense fog or heavy rain were not fully simulated in the current dataset.

Future work will focus on three directions:(1)expanding the dataset to include diverse weather conditions and crane types to enhance generalization;(2)investigating lightweight transformer-based backbones to capture long-range dependencies for better small object detection.

 doi:10.20944/preprints202512.2716.v1

5. Conclusions

To address the dual requirements of high precision and real-time performance for obstacle detection in complex and dynamic gantry crane operating scenarios, this study proposes an advanced detection framework based on an improved YOLOv5s model. The framework integrates a parameter-free SimAM loss function, and a joint image preprocessing pipeline comprising sharpening, denoising, and low-light enhancement. Validation conducted on a self-constructed dataset containing 1, 620 images across five object categories demonstrates that the proposed model achieves a mean Average Precision of 95. 2% with an inference latency of only 20. 1 ms, significantly outperforming the baseline YOLOv5s and other mainstream models such as SE and CBAM. Specifically, the SimAM module enhances critical feature extraction without increasing parameter overhead, while the EIOU loss function improves bounding box regression precision by optimizing aspect ratio differences and localization metrics. Furthermore, the preprocessing pipeline effectively mitigates environmental interference common in industrial settings. Ablation studies confirm the synergistic effects of these integrated modules. This model provides robust technical support for real-time industrial obstacle detection and facilitates the transition toward automated and safe crane operations. Future research will focus on model quantization, multimodal sensor fusion, and the optimization of visualized monitoring systems to further advance intelligent development in the crane industry.

Author Contributions: Formal analysis, B.L. and X.Z.; Investigation, L.S.; Data curation, B.L. and X.Z.; Writing—original draft, B.L., X.Z.; Writing—review and editing, L.S., L.Y. and K.L. All authors have read and agreed to the published version of the manuscript.

Funding: The project is supported by the 2022 Henan Province Industrial Research and Development Joint Fund Major Project (Project No. : 225101610072), 2024 Henan Province Industrial Research and Development Joint Fund Major Project (Project No. : 245101610033), 2025 Henan Province Science and Technology Key Project (Project No. : 252102411007), and 2025 National Administration for Market Regulation Science and Technology Program (Project No. : 2024MK080).

Data Availability Statement: The data presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li GJ. Key characteristics and control system general architecture of intelligent crane. J Mech Eng. 2020;56(24):254-68. DOI: 10. 3901/JME. 2020. 24. 254
2. Chen ZM, Li M, Shao XJ, Zhao ZC. Obstacle avoidance path planning for bridge crane based on improved RRT algorithm. *J Syst Simul*. 2021;33(8):1832-8. DOI: 10. 16182/j. issn1004731x. joss. 20-0272
3. Hongjie Z, Huimin O, Huan X. Neural network-based time optimal trajectory planning method for rotary cranes with obstacle avoidance. Mechanical Systems and Signal Processing. 2023;185. DOI: 10. 1016/j. Ymssp. 2022. 109777
4. Wa Z, He C, Haiyong C, Weipeng L. A Time Optimal Trajectory Planning Method for Double-Pendulum Crane Systems With Obstacle Avoidance. IEEE ACCESS. 2021;9:13022-30. DOI: 10. 1109/access. 2021. 3050258
5. Alkhaldi TM, Asiri MM, Alzahrani F, Sharif MM. Fusion of deep transfer learning models with Gannet optimisation algorithm for an advanced image captioning system for visual disabilities. Scientific Reports. 2025;15(1):40446-. DOI: 10. 1038/s41598-025-24171-9
6. Bing Z, Sinem G. Fine-Grained Visual Recognition in Mobile Augmented Reality for Technical Support. IEEE transactions on visualization and computer graphics. 2020;PP. DOI: 10. 1109/tvcg. 2020. 3023635

 doi:10.20944/preprints202512.2716.v1

7. Gao K, Chen L, Li Z, Wu Z. Automated Identification and Analysis of Cracks and Damage in Historical Buildings Using Advanced YOLO-Based Machine Vision Technology. Buildings. 2025;15(15):2675-. DOI: 10. 3390/buildings15152675
8. Wu Y, Liu M, Li J. Detection and Recognition of Visual Geons Based on Specific Object-of-Interest Imaging Technology. Sensors. 2025;25(10):3022-. DOI: 10. 3390/s25103022
9. Zou H, Yu XL, Lan T, Du Q, Jiang Y, Yuan H. Classification and recognition of black tea with different degrees of rolling based on machine vision technology and machine learning algorithms. Heliyon. 2025;11(14):e43862-e. DOI: 10. 1016/j. Heliyon. 2025. E43862
10. Akber MZ, Chan WK, Lee HH, Anwar GA. TPE-Optimized DNN with Attention Mechanism for Prediction of Tower Crane Payload Moving Conditions. Mathematics. 2024;12(19):3006-. DOI: 10. 3390/math12193006
11. Zhao Y, Zhang J, Zhang C. Deep-learning based autonomous-exploration for UAV navigation. Knowledge-Based Systems. 2024;297:111925-. DOI: 10. 1016/j. Knosys. 2024. 111925
12. Liu D. Research on lightweight monocular vision based end-side localization algorithm [Master's thesis]. Beijing: Beijing University of Posts and Telecommunications; 2024. DOI:2024. 10. 26969/d. cnki. gbydu. 2024. 000947
13. Gui DD. Research and implementation of traffic safety helmet wearing detection system based on deep learning [Master's thesis]. Nanjing: Southeast University; 2023. DOI:10. 27014/d. cnki. gdnau. 2023. 001155
14. Peng H, Shaochun M, Chenyang S, Zhengliang D. Emergency obstacle avoidance system of sugarcane basecutter based on improved YOLOv5s. Computers and Electronics in Agriculture. 2024;216:108468-. DOI: 10. 1016/j. Compag. 2023. 108468
15. Kim K, Kim K, Jeong S. Application of YOLO v5 and v8 for Recognition of Safety Risk Factors at Construction Sites. Sustainability. 2023;15(20):15179. DOI: 10. 3390/su152015179
16. Guoyan Y, Yingtong L, Ruoling D. An detection algorithm for golden pomfret based on improved YOLOv5 network. Signal, Image and Video Processing. 2022;17(5):1997-2004. DOI: 10. 1007/s11760-022-02412-y
17. Junzhou C, Kunkun J, Wenquan C, Zhihan L, Ronghui Z. A real-time and high-precision method for small traffic-signs recognition. Neural Computing and Applications. 2021;34(3):2233-45. DOI: 10. 1007/s00521-021-06526-1
18. Wenjie L, Lu W. Quantum image edge detection based on eight-direction Sobel operator for NEQR. Quantum Information Processing. 2022;21(5). DOI: 10. 1007/s11128-022-03527-4
19. Yuan S, Li X, Xia S, Qing X, Deng JD. Quantum color image edge detection algorithm based on Sobel operator. Quantum Information Processing. 2025;24(7):195-. DOI: 10. 1007/s11128-025-04819-1
20. Sun T, Xu J, Li Z, Wu Y. Two Non-Learning Systems for Profile-Extraction in Images Acquired from a near Infrared Camera, Underwater Environment, and Low-Light Condition. Applied Sciences. 2025;15(20):11289-. DOI: 10. 3390/app152011289
21. Yang H, Wang W, Wang Y, Wang P. Novel method for robust bilateral filtering point cloud denoising. Alexandria Engineering Journal. 2025;127:573-85. DOI: 10. 1016/j. Aej. 2025. 04. 099
22. Zhou Y, Zhang T, Li Z, Qiu J. Improved Space Object Detection Based on YOLO11. Aerospace. 2025;12(7):568-. DOI: 10. 3390/aerospace12070568
23. Yuan X, Wang Y, Li Y, Kang H, Chen Y, Yang B. Hierarchical flow learning for low-light image enhancement. Digital Communications and Networks. 2025;11(04):1157-71. DOI: 10. 1016/j. dcan. 2024. 11. 010
24. Yuanfu G, Puyun L, Xiaodong Z, Lifei Z, Guanzhou C, Kun Z, et al. Enlighten-GAN for Super Resolution Reconstruction in Mid-Resolution Remote Sensing Images. Remote Sensing. 2021;13(6):1104-. DOI: 10. 3390/rs13061104
25. Wu YL. Research on road obstacle detection and distance measurement algorithm based on YOLOv5 [Master's thesis]. Wuhu: Anhui Polytechnic University; 2023. DOI:10. 27763/d. cnki. gahgc. 2023. 000351
26. Peng R, Liao C, Pan W, Gou X, Zhang J, Lin Y. Improved YOLOv7 for small object detection in airports: Task-oriented feature learning with Gaussian Wasserstein loss and attention mechanisms. Neurocomputing. 2025;634:129844-. DOI: 10. 1016/j. Neucom. 2025. 129844
27. Dong Z. Vehicle Target Detection Using the Improved YOLOv5s Algorithm. Electronics. 2024;13(23):4672-. DOI: 10. 3390/electronics13234672

 doi:10.20944/preprints202512.2716.v1

28. Yang XJ, Zeng ZY. Dy-YOLO: an improved object detection algorithm for UAV aerial photography based on YOLOv5. *J Fujian Norm Univ (Nat Sci Ed)*. 2024;40(1):76-86. DOI: 10. 12046/j. issn. 1000-5277. 2024. 01. 009
29. Doong SH. Predicting postural risk level with computer vision and machine learning on multiple sources of images. Engineering Applications of Artificial Intelligence. 2025;143:109981-. DOI: 10. 1016/j. Engappai. 2024. 109981