A Mechanical Part Inspection Method Based on Improved YOLOv8

Haiwei Wu and Yisen Wang *

Posted Date: 8 July 2025

doi: 10.20944/preprints202507.0596.v1

Keywords: mechanical parts detection; YOLOv8; GRF_SPPF; CA attention mechanism; SIoU loss

 doi:10.20944/preprints202507.0596.v1

Article

A Mechanical Part Inspection Method Based on Improved YOLOv8

Haiwei Wu and Yisen Wang *

Jilin Agricultural University, Changchun City, Jilin Province, China

* Correspondence: 20230278@mails.jlau.edu.cn

Abstract

In the field of industrial visual inspection, traditional horizontal detection box approaches often encounter challenges such as inadequate feature extraction for small-scale components, localization inaccuracies, and reduced detection efficiency when applied to densely arranged micro-mechanical assemblies. This study presents a novel enhancement of the YOLOv8 framework, introducing an algorithm specifically designed to address the precision inspection requirements of industrial environments. Firstly, a globally receptive field-enhanced spatial pyramid pooling module (GRF-SPPF) is proposed, which significantly improves the model's ability to capture fine-grained features of micro-screw heads—such as cross-groove depth and hexagonal profiles—through multi-scale feature fusion. Secondly, a coordinate attention (CA) mechanism is integrated, combining spatial and channel attention via coordinate decomposition. This mechanism enhances the model's joint spatial-channel representation capability while simultaneously reducing the number of parameters and computational cost. Thirdly, to effectively manage rotation-sensitive targets, the conventional intersection over union (IoU) metric is replaced with an angle-constrained Skew IoU (SIoU) metric, which substantially improves convergence speed and spatial localization accuracy. Empirical validation using real-world production line datasets demonstrates that the optimized model, with a compact size of 6.4 MB, achieves a 1.5 percentage point increase in mean Average Precision at 0.5 IoU (mAP@0.5), reaching 89.6%. These results confirm the model's ability to meet the dual demands of high detection accuracy and real-time performance in industrial inspection scenarios.

Keywords: mechanical parts detection; YOLOv8; GRF_SPPF; CA attention mechanism; SIoU loss

1. Introduction

In the domains of industrial manufacturing and recycling, mechanical part recognition technology plays a pivotal role by providing critical data support for automated sorting systems and intelligent assembly lines. In recent years, significant research efforts have been directed toward the development of detection models that integrate machine vision with deep learning, with the objective of continuously improving the reliability and robustness of recognition systems. Nevertheless, industrial environments often present practical challenges such as disordered part placement, imbalanced sample category distribution, and complex background interference. These factors contribute to several persistent limitations in existing detection models, including structural complexity, suboptimal accuracy, and limited algorithmic generalizability. To overcome these technical bottlenecks, the effective integration of advanced object detection algorithms into automated inspection systems has emerged as a focal point of current research, aiming to enhance both detection performance and system adaptability in real-world industrial applications.

As deep learning object detection continues to evolve, current detectors are divided into two main types: two-stage detection and single-stage detection [1].

Among object detection approaches, two-stage detection frameworks operate by first generating candidate regions—typically numbering in the hundreds to thousands—that are likely to contain objects. This preliminary step significantly reduces redundant computations in subsequent

 doi:10.20944/preprints202507.0596.v1

processing. In the second stage, features are extracted from each candidate region to predict object categories and refine bounding box positions. Representative models following this paradigm include RCNN [2,3], Fast R-CNN [4], Faster R-CNN [5]. In contrast, single-stage object detection is a more efficient approach within deep learning, characterized by its direct prediction of both object location and category in a single forward pass of the network, eliminating the need for region proposal generation. This methodology is recognized for its speed and suitability for real-time applications. Notable models in this category include the YOLO [6] series, SSD [7], and RetinaNet [8]. Among these, the YOLO series has emerged as the most widely adopted single-stage detection model, gaining substantial attention for its balance of real-time performance and detection accuracy. Recent advancements include the work of Pu et al. [9], who optimized the Transformer-based edge detection model (EDTER) by incorporating deformable convolution to enhance edge localization accuracy. Additionally, they integrated global self-attention mechanisms to improve semantic feature fusion, further advancing the capabilities of edge-aware object detection systems. However, in practical deployment, the high computational complexity restricts real-time performance, making it challenging to balance detection speed and accuracy in low-resource environments (such as mobile devices), thereby constraining its generalization capability for industrial applications. Building upon YOLOv5, Zhu et al. [10] introduced a Transformer Prediction Head (TPH) after the backbone network to develop the TPH-YOLOv5 detection model. This approach improves the capture of small-target features in complex scenes through multi-scale self-attention; however, the significant increase in inference time resulting from the multi-head attention mechanism leads to substantial memory consumption. This creates a trade-off between frame rate and accuracy in real-time detection of high-resolution images, ultimately limiting its feasibility for deployment on mobile devices. Bochkovskiy et al. [11], based on the YOLOv3 architecture, introduced the CSPDarknet53 backbone network, along with spatial pyramid pooling (SPP) and the path aggregation network (PANet), thereby significantly enhancing detection speed and accuracy. However, the substantial number of parameters (approximately 60 million) leads to high memory consumption, resulting in inference latencies exceeding 200 ms on embedded devices (such as the Raspberry Pi 4B), making it challenging to meet the real-time performance requirements of industrial applications. S. Sabour et al. proposed a dynamic routing mechanism based on [12] (Capsule Network) by iteratively propagating feature importance weights from the bottom up, which achieved dynamic path selection for multi-scale features. The robustness of this method was validated on the MNIST dataset for complex structures (such as overlapping components); however, due to the high computational cost of iterations—resulting in an approximate 40% increase in inference time per image—it remains challenging to deploy for real-time edge detection tasks.

2. YOLOv8 Network Model

The YOLO series of network models represents an optimal choice for industrial part inspection, offering a well-balanced trade-off between speed and accuracy. YOLOv8 [13,14], in particular, surpasses its predecessors in detection performance, usability, and deployment flexibility through architectural innovations and algorithmic optimizations. These advancements enable it to more effectively meet the core requirements of high real-time responsiveness, robust performance, and low-cost deployment in industrial environments.

YOLOv8 represents a more advanced iteration within the YOLO series, demonstrating notable improvements over earlier versions such as YOLOv7, YOLOv5, and YOLOv4 in terms of algorithmic efficiency, detection performance, and lightweight design. It successfully maintains high operational efficiency while simultaneously enhancing detection accuracy [15]. Nevertheless, its current loss function does not fully exploit global perception capabilities, and the employed feature fusion strategy constrains its effectiveness in detecting small-scale targets [16]. The core architecture of YOLOv8 consists of three parts: Backbone (the backbone network), neck (the neck network), and head (the detection head). Below is an overview of the development process and key improvements of these three components in the YOLO series: 1. Backbone (backbone network): YOLOv8 removes the

 doi:10.20944/preprints202507.0596.v1

focus module from YOLOv5 and replaces it with a more efficient convolutional layer to simplify the structure; it optimizes the CSP module to C2f (cross-stage partial residual connection), enhances gradient flow and reduces the number of parameters. It also retains spatial pyramid pooling (SPPF) to accelerate multi-scale feature fusion. 2. Neck (neck network): YOLOv8 adopts the PANet structure and further simplifies the neck structure compared with YOLOv7, integrating it into the first half of the head. This is achieved through upsampling (Upsample) and Concat operations for multi-scale feature fusion and introduces a dynamic path selection mechanism to increase flexibility. 3. Head (Detection Head): The head is fully transitioned to the anchor-free design, directly predicting the target center and width-height, reducing complexity; it adopts a decoupled head (decoupled head), separating classification from regression tasks to minimize task conflicts. Additionally, dynamic label allocation (task-aligned assignment) is introduced, dynamically allocating positive and negative samples on the basis of task difficulty and enhancing training efficiency. The YOLOv8 network model diagram is shown in Figure 1.

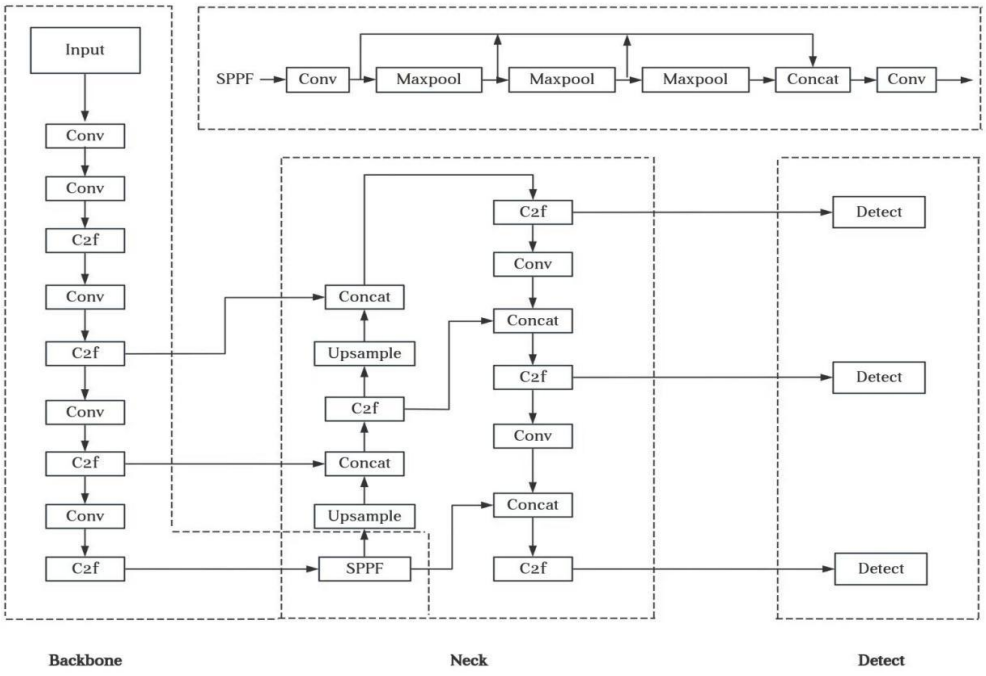


Figure 1. YOLOv8 network model diagram.

3. YOLOv8 Improves the Network Model

For the identification problem of high–density arrangements of micromechanical parts, the traditional horizontal detection frame method often encounters problems such as insufficient feature extraction of small–sized parts, positioning deviation and low detection efficiency. The structure of YOLOv8 is depicted in Figure 2 [17].

 doi:10.20944/preprints202507.0596.v1

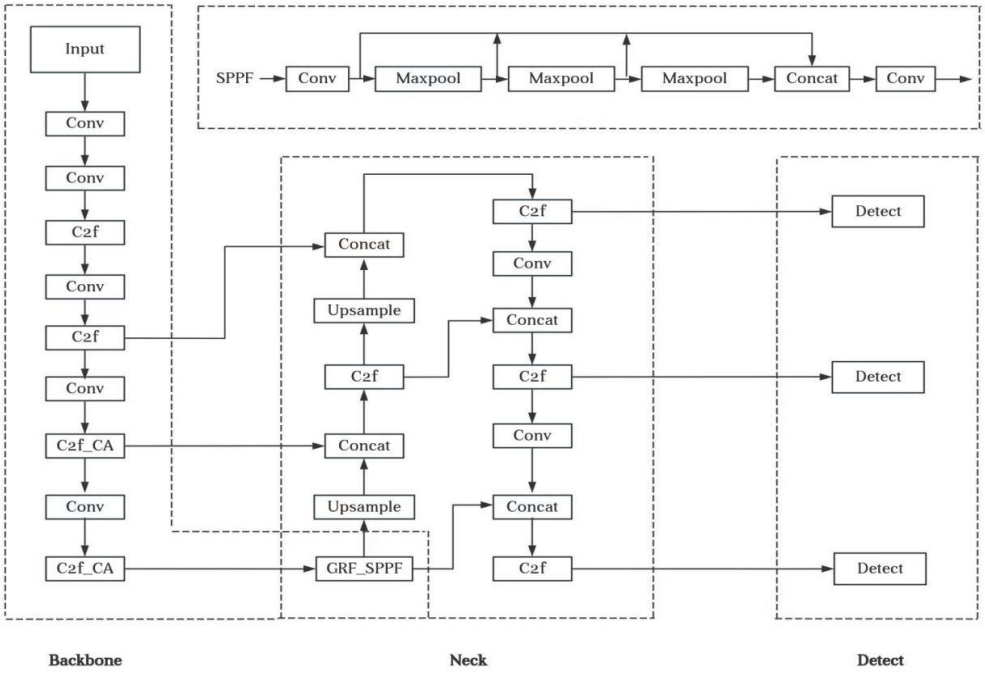


Figure 2. YOLOv8 Network improvement model diagram.

(1) The spatial pyramid pooling is improved, and the SPPF module is substituted by the GRF-SPPF [18] module, which can effectively improve the model's fine-grained perception ability for the microscrew head features.

(2) The CA [19] attention mechanism is introduced into the backbone (main network) part to better adapt based on different tasks and input data.

(3) The SIoU [20] metric with angular constraints is employed in place of the traditional intersection-over-union calculation. This substitution markedly enhances the model's convergence speed and spatial localization accuracy.

3.1. GRF-SPPF

In the field of object detection and image recognition, the spatial pyramid pooling fast (SPPF) module is a staple in mainstream algorithms, including those in the YOLO series, due to its efficiency in multi-scale feature extraction. However, traditional SPPF modules exhibit limitations in capturing global contextual information. To address this, the present study introduces the GRF-SPPF module as an innovative enhancement. By incorporating a dual-path mechanism comprising global average pooling (GAP) and global max pooling (GMP), a hybrid pooling architecture is constructed, endowing the model with stronger feature representation capabilities. The structural configuration of the GRF-SPPF module is depicted in Figure 3.

The core of this technology lies in the construction of a three-level feature fusion system. First, the original SPPF module extracts local features from various receptive fields through parallel max pooling layers, thereby forming a multi-scale pyramid structure. Second, a newly introduced global average pooling layer captures the image's overall statistical features by compressing spatial dimensions, providing the network with global semantic context. Simultaneously, a global max pooling layer enhances the saliency of target edges and texture features. Following these processes, the three streams of features are concatenated along the channel dimension. A series of 1×1 convolution operations are then applied to facilitate cross-channel information interaction and to perform dimensionality reduction, effectively retaining salient features while maintaining control over computational complexity.

 doi:10.20944/preprints202507.0596.v1

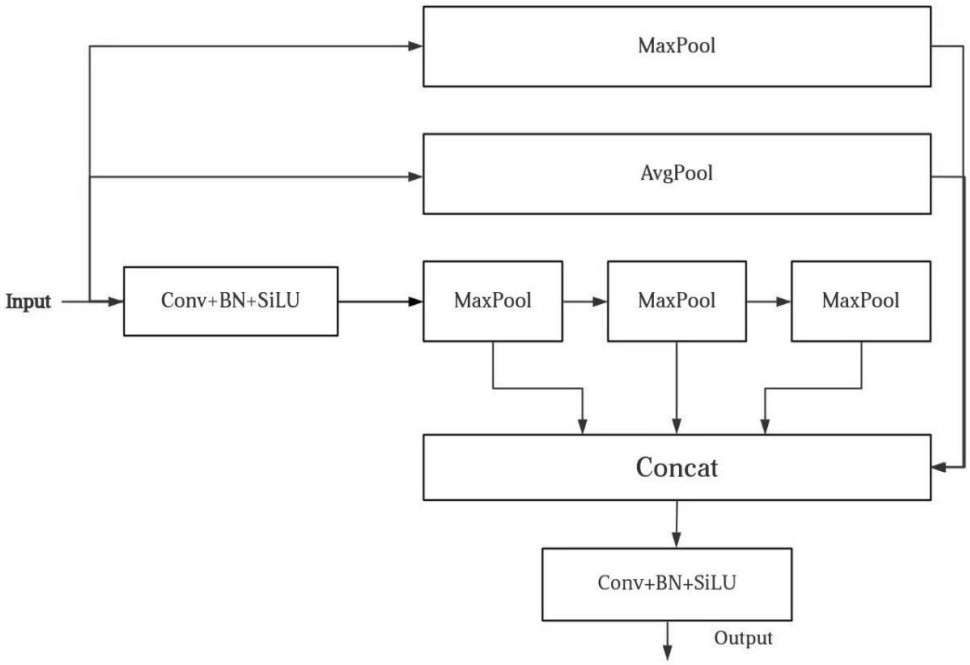


Figure 3. GRF-SPPF Network structure diagram.

The advantage of this approach lies in its ability to achieve complementary fusion of local and global features along the spatial dimension, enabling the network to simultaneously acquire fine-grained detail perception and a comprehensive understanding of the image context. Along the channel dimension, max pooling emphasizes salient features, with the feature vectors derived from the two max pooling pathways exhibiting distinct statistical characteristics. Meanwhile, average pooling captures the overall trend of feature distribution. The synergistic interaction between these pooling strategies contributes to an enhancement in the mAP@0.5 performance metric. In terms of computational efficiency, the use of 1 × 1 convolution layers ensures real-time inference capability, while incurring only a minimal increase in the number of parameters.

Compared with existing methods, this solution preserves the original advantages of the SPPF module while effectively addressing two major limitations inherent in traditional approaches: first, the loss of global context information caused by localized pooling operations; and second, the feature representation bias introduced by reliance on a single pooling strategy.

Compared with classic feature pyramid networks (FPNs) and dilated spatial pyramid pooling (ASPP) modules, the improved SPPF shows improvements in inference speed and context information capture.

3.2. Coordinate Attention (CA) Attention Mechanism

Coordinate attention (CA) is a new attention mechanism proposed in 2021. CA is an efficient spatial-channel collaborative attention mechanism whose core innovation lies in decomposing two-dimensional spatial features into horizontal and vertical one-dimensional coordinate directions for independent analysis. The structure of the CA attention mechanism is depicted in Figure 4.

 doi:10.20944/preprints202507.0596.v1

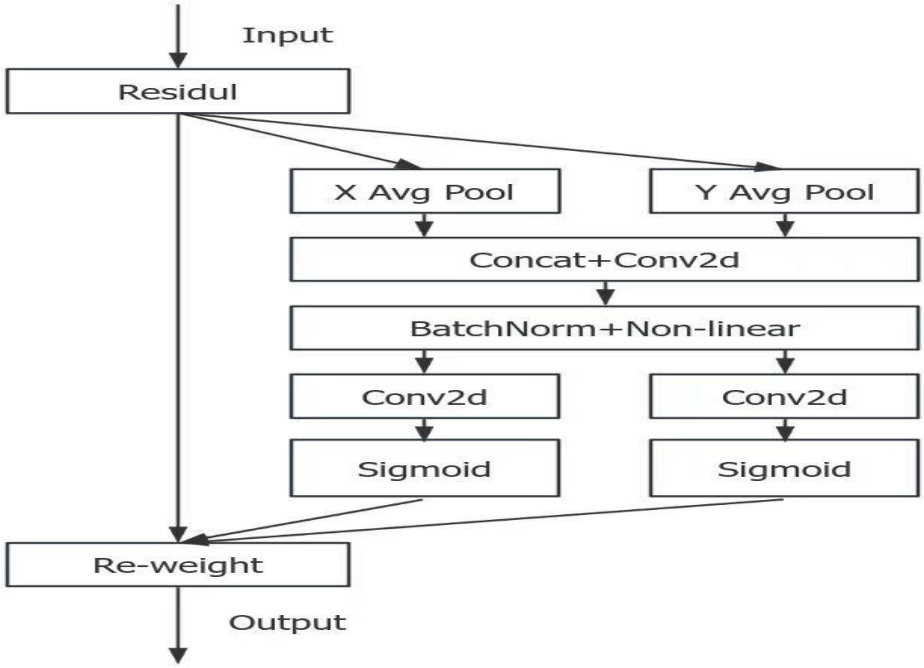


Figure 4. CA attention mechanism Network structure diagram.

The workflow of this mechanism is divided into two major stages. First, coordinate information embedding (Coordinate Information Embedding) performs global average pooling on the input feature map in both the horizontal and vertical directions to obtain feature vectors in both directions. Specifically, this involves extracting horizontal information (performing 1D average pooling along the width direction (W)), resulting in an output size of C×H×1. The merging formula is Formula (1):

$$z_c^h(h) = \frac{1}{W}\sum_{0\le w<W} x_c\,(h, w) \qquad (1)$$

In addition, the vertical information is extracted (1D average pooling along the height direction (H)), and the output size becomes C×1×W. By employing this step, the model can capture the spatial location information of the feature map. The merging formula is Formula (2):

$$z_c^w(w) = \frac{1}{H}\sum_{0\le h<H} x_c\,(h, w) \qquad (2)$$

Then, coordinate attention generation (Coordinate Attention Generation):the specific operation is to first splice the feature vectors in the horizontal and vertical directions (Z^h and $Z^w$are spliced along the spatial dimension, after obtaining $Z^{cat} \in R^{C\times(H+W)\times1}$, and then through a convolutional layer with nonlinear activation, generate intermediate features $F \in R^{C/r\times(H+W)\times1}$) The formula is Formula (3):

$$F = \delta\left(\mathrm{Conv1D}_{1\times1}(Z^{cat})\right) \qquad (3)$$

Then, F is decomposed into horizontal component $F^h \in R^{C/r\times H\times1}$ and vertical components $F^w \in R^{C/r\times W\times1}$, and the channel dimension is restored through convolution. The horizontal attention weight formula is Formula (4):

$$A^h = \sigma\left(\mathrm{Conv1D}_{1\times1}(F^h)\right) \in R^{C\times H\times1} \qquad (4)$$

The vertical attention weight is Formula (5):

$$A^w = \sigma\left(\mathrm{Conv1D}_{1\times1}(F^w)\right) \in R^{C\times1\times W} \qquad (5)$$

Finally, the original feature map X is multiplied point by point with the attention weight. The formula is Formula (6):

$$X_{out} = X \odot A^h \odot A^w \qquad (6)$$

final output.$X_{out} \in R^{C\times H\times W}$

 doi:10.20944/preprints202507.0596.v1

Compared with traditional attention mechanisms, CA breaks through by ① retaining absolute position information through coordinate decomposition, enhancing the ability to locate small targets; ② using one-dimensional convolution to reduce computational complexity, reducing parameters compared with conventional two-dimensional attention; and ③ establishing cross-channel spatial associations, effectively distinguishing dense targets. In industrial inspection scenarios, these changes can accurately identify the outlines of tiny parts, suppress background interference, and maintain stable detection capabilities for occluded targets, making them key technologies that improve the performance of YOLO series models in small-object detection tasks.

3.3. Loss Function SIoU

In YOLOv8, target box regression employs the CIoU loss function [21]. However, CIoU does not explicitly incorporate angular deviation between predicted and ground truth bounding boxes. As a result, it can lead to substantial localization errors when detecting rotated targets, thereby diminishing the optimization effectiveness of the regression process [22]. In object detection tasks, the localization accuracy of boundary boxes (Bounding Box) directly impacts the model's detection effectiveness. Traditional loss functions such as the IoU [23] optimize localization by calculating the intersection-over-union ratio between the predicted and real boxes. However, in complex scenarios (such as object rotation or dense arrangements), these functions often face issues such as slow convergence and large positioning errors. To address the rotational sensitivity of high-density microparticles in industrial visual inspection, the Synthetic Intersection over Union (SIoU) loss function incorporates angular deviation into the loss calculation, uses angle constraints to optimize the rotation direction of the predicted box, and then combines distance and shape similarity for comprehensive adjustment. The core idea of the SIoU is that its formula consists of three parts: the vector angle diagram between the real and predicted boxes is depicted in Figure 5:

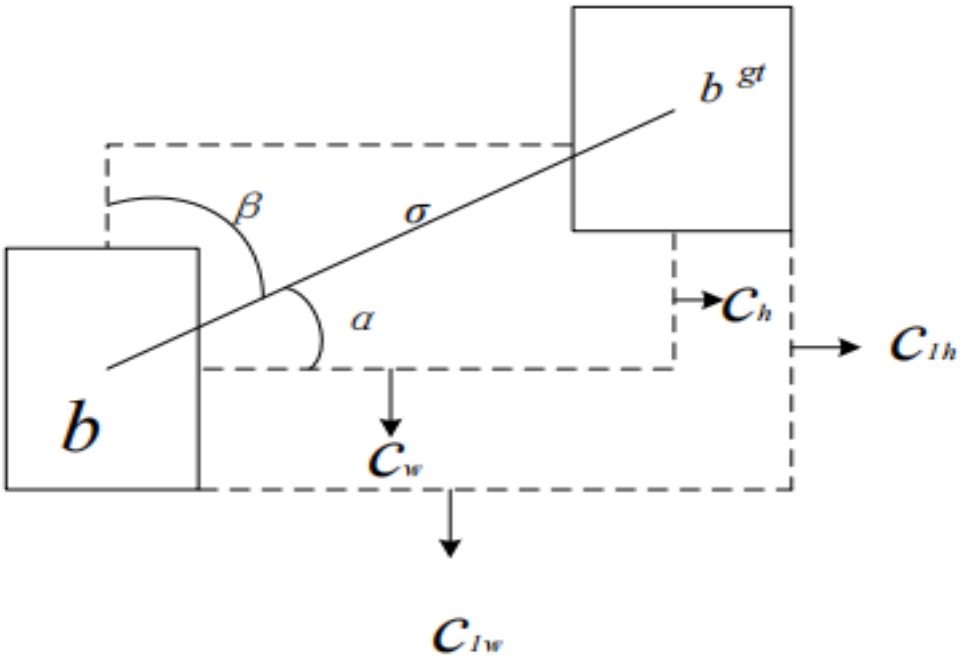


Figure 5. The scheme for calculation of angle cost contribution into the loss function.

The first is angle loss (Angle Cost). We first define the center coordinate of the real box as $\left(b_{c_x}^{gt}, b_{c_y}^{gt}\right)$, $\left(b_{c_x}, b_{c_y}\right)$ are the center coordinate of the predicted box.$c_h$is the height difference between the center points of the real box and the predicted box in the picture, The formula is Formula (7):

 doi:10.20944/preprints202507.0596.v1

$$c_h = \max\left(b_{c_y}^{gt}, b_{c_y}\right) - \min\left(b_{cy_y}^{gt}, b_{c_y}\right) \quad (7)$$

The distance between the center point of the real box and the predicted box is σ in the figure. The formula is Formula (8):

$$\sigma = \sqrt{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2} \quad (8)$$

Therefore, the angular $\frac{c_h}{\sigma} = \sin(\alpha)$.Then the angular loss is Formula (9):

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin\left(\frac{c_h}{\sigma}\right) - \frac{\pi}{4}\right) = 1 - 2 \cdot \sin^2\left(\theta - \frac{\pi}{4}\right) \quad (9)$$

The second is distance loss (Distance Cost) (Note: Here $(c1_h,\ c1_w)$ is the width and height of the minimum bounding rectangle of the real frame and the predicted frame),then, distance loss is Formula (10):

$$\Delta = \sum_{t=x,y}\left(1 - e^{-\gamma\rho_t}\right) = 2 - e^{-\gamma\rho_x} - e^{-\gamma\rho_y} \quad (10)$$

among, $\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2$ $\rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2$ $\gamma = 2 - \Lambda$

The third is shape loss (Shape Cost),$(w, h)$and(w^{gt}, h^{gt}) are the width and height of the predicted frame and the real frame respectively.θ controls the degree of attention to shape loss, so the shape loss is Formula (11):

$$\Omega = \sum_{t=w,h}\left(1 - e^{-w_t}\right)^\theta = \left(1 - e^{-w_w}\right)^\theta + \left(1 - e^{-w_h}\right)^\theta \quad (11)$$

among, $w_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} w_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}$.

In summary, the final SIoU loss function is Formula (12):

$$\mathcal{L}_{SIoU} = 1 - IoU + \frac{\alpha\Lambda + \beta\Delta + \gamma\Omega}{3} \quad (12)$$

Among them, α, β, γ are equilibrium coefficients, $IoU = \frac{b^{gt} \cap b}{b^{gt} \cup b}$ is used to adjust the degree of contribution of different losses.

The SIoU loss function addresses the limitations of traditional loss functions in object detection involving rotated targets by incorporating angular constraints, dynamic distance penalties, and shape similarity optimization. Its design rationale aligns closely with the core demands of industrial visual inspection, such as high-precision localization and lightweight deployment. As a result, it offers an effective solution for applications that include micro-part recognition and real-time quality inspection on production lines. In the future, the SIoU is expected to exhibit even greater potential in complex industrial environments.

4. Experimental Results and Analysis

4.1. Experimental Environment and Dataset

This experiment is carried out on a cloud computing platform, and the related hardware and software resource allocations are shown in Table 1.

Table 1. Hardware and software resource allocation.

Configure	Parameter
GPU	RTX4090 (24GB) 1 elevation configuration
CPU	16vCPUIntel(R)Xeon(R)Platnum8352VCPU@210GHZ
internal storage:	120GB
Hard disk system disk:	30GB
PyTorch	1.11.0
python	3.8
cuda	11.2

 doi:10.20944/preprints202507.0596.v1

The hyperparameter configuration of the network training in this experiment is shown in Table 2.

Table 2. Hyperparameter configuration.

Name	Set the number
epochs	120
batch	32
imgsz	640
optimizer	SGD
patience	50
lr0	0.01
close_mosaic	0
weight_decay	0.0005

The experimental dataset was acquired using a German Basler SCA640-70fm industrial camera, with image dimensions set to 640 × 640 pixels, comprising a total of 3,931 frames. The dataset encompasses six categories of industrial components: bearings, bolts, flanges, gears, nuts, and springs. These parts exhibit texture colors highly similar to the material frame background, contributing to a visually complex environment. Sample annotation was performed using the LabelImg annotation tool, with the resulting labels saved in YOLO format as .txt files. The annotated data were divided into training, evaluation, and prediction sets at an 8:1:1 ratio, some representative sample effects is depicted in Figure 6.

Figure 6. Three categories of parts data sets.

4.2. Evaluation Indicators

To measure the performance of the model, the following general target detection evaluation indicators are selected: accuracy (P), recall rate (R), and average precision mean (mAP@0.5) [24]:

4.2.1. Precision (Accuracy) and Recall (Recall Rate)

Precision (accuracy): The proportion of the results predicted by the model as positive samples that are actually positive samples. The formula is Formula (13):

$$P = \frac{TP}{TP+FP} \times 100\% \quad (13)$$

Among them, true positive (TP) is the true positive example, and false positive (FP) is the false positive example.

Recall (recall rate): indicates the proportion of all actual positive samples that are correctly predicted as positive samples by the model. The formula is Formula (14):

 doi:10.20944/preprints202507.0596.v1

$$R = \frac{TP}{TP+FN} \times 100\% \quad (14)$$

Among them, false negative (FN) is a false negative.

4.2.2. AP (Average Precision, Average Accuracy)

AP is the average precision calculated at different recall thresholds. The area under the P-R curve, obtained by plotting precision-recall (P-R) curves, is used as the value of the AP. In other words, the AP measures the model's detection capability in a specific category. The formula is Formula (15):

$$AP = \int_0^1 P(r)dr \quad (15)$$

P-R curve: with the recall rate on the horizontal axis and accuracy on the vertical axis, the curve shows the change in model accuracy under different recall rates. The higher the AP value is, the better the model's accuracy performance is under different recall rates.

4.2.3. MAP@0.5 (Mean Average Precision, Mean Average Accuracy)

mAP@0.5 is the average AP in multiple categories and is used to measure the detection performance of the model on the whole dataset. The specific process first calculates the AP of each category and then takes the average of all categories to obtain mAP@0.5. The formula is Formula (16):

$$mAP = \frac{\sum_{i=0}^{n} AP(i)}{n} \quad (16)$$

4.3. Experiments and Comparison

4.3.1. Ablation Experiment

To evaluate the effectiveness of algorithm improvements systematically, this study designed ablation experiments (ablation studies) based on the YOLOv8 framework and quantitatively analyzed the impact of each component on model performance by gradually introducing innovative modules. The results are shown in Table 3, with seven experimental groups designed in total. The first group is the base model YOLOv8 (baseline mAP@0.5 at 88.1%). The second group introduces the GRF-SPPF alone, which improves the mAP@0.5 by 0.6 percentage points compared with the baseline, enhancing the fine-grained perception of small targets through multi-scale feature fusion. The third group introduces the CA mechanism alone, increasing mAP@0.5 by 0.5 percentage points compared with the baseline, making its lightweight design particularly suitable for industrial scenarios with limited computing resources. The fourth experimental group employs the SIoU metric in place of the traditional IoU, resulting in a 0.4 percentage point improvement in mAP@0.5. This enhancement is particularly significant for detecting overlapping instances among densely arranged components. The first four experimental groups were designed to independently evaluate the effectiveness of the individual improvement methods proposed in this study [25].

Table 3. shows the ablation results.

Model	GRFSPPF	CA	SIOU	mAP@0.5/%
YOLOv8	×	×	×	88.1
A	√	×	×	88.7
B	×	√	×	88.6
C	×	×	√	88.5
D	√	√	×	89.2
E	√	×	√	89.1
Proposed method	√	√	√	89.6

To evaluate the impact of different combinations of improvements on algorithm performance, a dual-module synergy strategy was employed. Specifically, Group 5 (Model D, mAP@0.5 = 89.2%)

 doi:10.20944/preprints202507.0596.v1

integrates the GRF-SPPF and CA modules, combining multi-scale feature enhancement with attention mechanisms. This configuration yielded a 0.5 percentage point improvement in mAP@0.5 compared to the standalone CA module. GRF-SPPF enables the network to perceive both fine-grained details and global contextual information through the complementary fusion of local and global features. When combined with the CA mechanism, which emphasizes critical regions, the model demonstrates enhanced robustness in detecting micro-part contours, such as hexagonal edges. Group 6 (Model E, mAP@0.5 = 89.1%) combines GRF-SPPF with the SIoU loss function, uniting multi-scale feature extraction with angle-constrained localization optimization. This pairing achieves an mAP@0.5 of 89.1%. While GRF-SPPF addresses SIoU's limitations in capturing complex texture features, SIoU enhances spatial localization accuracy, establishing a mutually complementary relationship. The final proposed model integrates all three improvement modules—GRF-SPPF, CA, and SIoU—achieving a notable mAP@0.5 of 89.6%, representing a 1.5 percentage point increase over the baseline. In practical industrial applications, the model retains a compact size of 6.4 MB and supports real-time detection, effectively addressing the challenges of high-density, multi-angle micro-part inspection tasks on production lines.

To demonstrate the effectiveness of each improvement module proposed in this study [26], ablation experiments were conducted to systematically evaluate both the individual contributions and the synergistic effects of the modules. The results confirm that the final algorithm achieves an optimal balance between detection accuracy and computational efficiency, offering a reliable and practical solution for precision inspection in industrial settings. Future research can further explore automated combination optimization of modules or integrate more efficient lightweight designs (such as neural network architecture search) to adapt to more complex industrial environment requirements.

4.3.2. Comparative Experiments

To verify the effectiveness of the improved algorithm in terms of detection performance, it is compared with current mainstream target detection algorithms, such as the Faster R-CNN, SSD, and YOLO series algorithms. The comparison experiment [27] is carried out on the same dataset as the improved algorithm. The detection results are shown in Table 4.

Table 4. Comparison experiment of custom-made datasets.

Model	Params/M	GFLOPs/G	FPS/(f·s-1)	mAP@0.5/%
Faster R-CNN	41.5	207.3	12	89.2
SSD512	26.8	99.6	45	85.7
YOLOv3-608	61.5	154.7	35	87.4
YOLOv5n	46.5	109.1	95	88.9
YOLOv7n	37.2	104.7	105	89.1
YOLOv8s	3.1	8.2	120	88.1
Ours	6.4	9.3	105	89.6

From the perspective of algorithm parameter volume and computational cost, the improved YOLOv8 model has a parameter size of 6.4M. Although the parameter count of the proposed model represents an increase compared to the lightweight YOLOv8s network, it remains only one-tenth that of the traditional YOLOv3-608 model, and is substantially lower than that of Faster R-CNN and SSD512. This design strategy also confers notable advantages in managing computational complexity. The improved model achieves a computational cost of 9.3GFLOPs, a significant reduction compared to the 154.7 GFLOPs required by YOLOv3-608. Furthermore, it maintains a minimal increase of only 13.4% over the 8.2GFLOPs of YOLOv8s, clearly demonstrating the efficiency and compactness achieved through targeted algorithmic optimization.

In terms of detection accuracy, the improved model achieved an mAP@0.5 of 89.6%, surpassing the highest-performing models in the comparison group, including YOLOv7n and Faster R-CNN.

 doi:10.20944/preprints202507.0596.v1

Remarkably, the model maintained high precision while achieving a real-time detection speed of 105 FPS—nearly eight times faster than Faster R-CNN's 12 FPS—and on par with YOLOv7n. This demonstrates that the proposed improvements effectively address the common trade-off between precision and speed typically observed in traditional optimization strategies. Compared to YOLOv8s, which has a similar parameter count and achieved an mAP@0.5 of 88.1%, the improved model shows a 1.5 percentage point increase in accuracy. This gain highlights the enhanced feature representation capacity enabled by the refined network architecture.

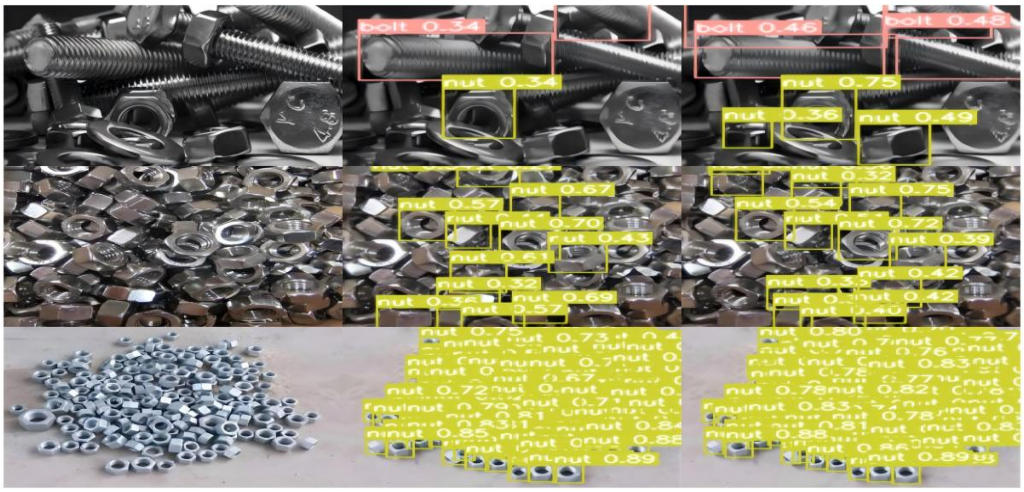
From the perspective of algorithm update trends, this study reveals the optimization potential of YOLO series models. Compared with earlier versions YOLOv3-608 and YOLOv5n, the improved model achieves a significant accuracy increase while reducing the number of parameters by more than 90%, demonstrating that the lightweight design does not necessarily come at the cost of detection performance. The comparison with the latest version YOLOv8s is even more enlightening: when the number of parameters increases by about 106%, the improved model maintains a relative speed of 87.5% while achieving an accuracy improvement of 1.5%. Notably, the improved model demonstrates unique advantages in balancing computational resource usage and detection efficiency. With a computational cost of only 9.3 GFLOPs, which is just 9.3% of SSD512's, it achieves a 3.9 percentage point accuracy improvement; compared with Faster R-CNN at the same precision level, its computational efficiency (GFLOPs/FPS) improves by two orders of magnitude.

Overall, this study successfully overcomes the bottleneck of improving the accuracy of lightweight models through innovative network structure adjustments. The improved YOLOv8 outperforms mainstream detection models in terms of accuracy at the cost of moderately increased parameters while maintaining real-time detection speed. This achievement not only validates the effectiveness of the proposed algorithmic optimizations but also introduces a novel solution for object detection that successfully balances accuracy and efficiency. It offers substantial practical value for real-world applications that demand high real-time performance, such as the automated recognition and sorting of recyclable components.

4.4. Comparison of Experimental Results

To evaluate the effectiveness of the algorithm in real-world applications, this study utilized a dataset comprising mechanically recycled and sorted parts collected from actual factory environments. Visual comparisons were conducted to assess the detection performance of the YOLOv8 model before and after the proposed improvements. Figure 7b illustrates the detection results of the original model, while Figure 7c presents the performance of the enhanced model. The comparative analysis reveals several notable advantages of the improved model. In industrial settings characterized by uneven lighting and complex textures, the enhanced algorithm demonstrates superior capability in identifying micro-particles that were previously missed. Additionally, in scenarios involving densely distributed objects, the accuracy of bounding box localization is significantly increased, and the false positive rate for overlapping targets is markedly reduced. These performance gains are primarily attributed to the integration of an extended receptive field module within the network architecture and the optimization of multi-scale feature fusion mechanisms. Together, these enhancements support the model's capacity to simultaneously capture fine-grained local details and global semantic context. The experimental detection outcomes are depicted in Figure 7.

 doi:10.20944/preprints202507.0596.v1



(a) (b) (c)

Figure 7. Part of the component instance detection comparison.

5. Conclusions

This study addresses the challenge of inspecting densely arranged micromechanical parts in industrial settings by proposing an innovative improved algorithm based on the YOLOv8 framework. The aim is to overcome performance bottlenecks in traditional methods when dealing with tiny target feature extraction, rotation-sensitive localization, and complex background interference. First, a global receptive field-enhanced spatial pyramid pooling module (GRF-SPPF) is designed.

By introducing additional dual-path mechanisms—Global Average Pooling (GAP) and Global Max Pooling (GMP)—this module enhances the model's capability to perceive fine-grained features of micro-components, effectively mitigating missed detections of small targets. Furthermore, a lightweight CA mechanism is incorporated, which fuses spatial and channel dimension features through coordinate decomposition strategies. This reduces the parameter size while improving the model's accuracy in distinguishing part positions and shapes, making the model lighter and more suitable for practical applications and deployment [28]. To address the localization demands of rotation-sensitive targets in industrial environments, the traditional IoU metric is replaced with a combined SIoU metric. By jointly optimizing angular deviation and spatial distance, this approach significantly improves convergence efficiency and localization robustness. Comparative analyses and ablation experiments have validated the effectiveness of these enhancements. Experiments on actual production line datasets show that the improved model, while maintaining its lightweight nature (6.4 MB), has an average detection accuracy (mAP@0.5) of 89%. 6, fully meeting the dual requirements of high precision and real-time industrial testing. Future research will focus on enhancing robustness under dynamic environmental conditions, designing adaptive data balancing strategies, and advancing edge computing deployment, thereby supporting the broader application and scalability of intelligent industrial inspection technologies.

Author Contributions: The following statements should be used "Conceptualization, Yisen Wang. and Haiwei Wu.; methodology, Yisen Wang.; software, Yisen Wang.; validation, Yisen Wang., Haiwei Wu..; formal analysis, Yisen Wang.; investigation, Yisen Wang.; resources, Yisen Wang.; data curation, Yisen Wang.; writing—original draft preparation, Yisen Wang.; writing—review and editing, Yisen Wang.; visualization, Yisen Wang.; supervision, Yisen Wang.; project administration, Yisen Wang.; funding acquisition, Haiwei Wu. All authors have read and agreed to the published version of the manuscript." Please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

 doi:10.20944/preprints202507.0596.v1

Data Availability Statement: Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

Acknowledgments: The authors wish to express sincere gratitude to Professor Haiwei Wu for his invaluable guidance, critical insights, and unwavering support throughout this research. His expertise and mentorship were instrumental in shaping the direction of this work. We thank our colleagues at College of Engineering and Technology, Jilin Agricultural University for stimulating discussions and constructive feedback. Special appreciation is extended to the anonymous reviewers of Natural Computing for their thorough evaluations and thoughtful suggestions, which significantly enhanced the rigor and clarity of this manuscript. Finally, we acknowledge the academic environment and resources provided by Jilin Agricultural University that facilitated this study.

Conflicts of Interest: All authors disclosed no relevant relationships.

References

1. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; pp. 7263-7271.
2. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016; pp. 779-788.
3. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014; pp. 580-587.
4. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015; pp. 1440-1448.
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137-1149.
6. Shafiee, M.J.; Chywl, B.; Li, F.; Wong, A. Fast YOLO: A fast you only look once system for real-time embedded object detection in video. *J. Comput. Vis. Imaging Syst.* **2017**, *3*. doi: 10.15353/vsnl.v3i1.171.
7. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Computer Vision – ECCV 2016, Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer, 2016; 9905 9905, pp. 21-37.
8. Abdo, H.; Amin, K.M.; Hamad, A.M. Fall detection based on RetinaNet and MobileNet convolutional neural networks. In 2020 15th International Conference on Computer Engineering and Systems (ICCES), IEEE: Cairo, Egypt, 2020; pp. 1-7.
9. Pu, M.; Huang, Y.; Liu, Y.; Guan, Q.; Ling, H. EDTER: Edge detection with transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022; pp. 1402-1412.
10. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021; pp. 2778-2788.
11. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**. doi: 10.48550/arXiv.2004.10934.
12. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **2017**. doi: 10.48550/arXiv.1710.09829.
13. Varghese, R.; Sambath, M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), IEEE: Chennai, India, 2024; pp. 1-6.
14. Vijayakumar, A.; Vairavasundaram, S. Yolo-based object detection models: A review and its applications. *Multimed. Tools Appl.* **2024**, *83*, 83535-83574.
15. Wang, C.; Wang, C.; Wang, L.; Li, Y.; Lan, Y. Real-time tracking based on improved YOLOv5 detection in orchard environment for dragon fruit. *J. ASABE* **2023**, *66*, 1109-1124.

 doi:10.20944/preprints202507.0596.v1

16. Zhang, P.; Liu, Y. A small target detection algorithm based on improved YOLOv5 in aerial image. *PeerJ Comput. Sci.* **2024**, *10*, e2007.
17. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE: Rhodes Island, Greece, 2023; pp. 1-5.
18. Xie, G.; Xu, Z.; Lin, Z.; Liao, X.; Zhou, T. GRFS-YOLOv8: an efficient traffic sign detection algorithm based on multiscale features and enhanced path aggregation. *Signal Image Video Process.* **2024**, *18*, 5519-5534.
19. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021; pp. 13713-13722.
20. Chen, Z.; Liu, C.; Filaretov, V.F.; Yukhimets, D.A. Multi-scale ship detection algorithm based on YOLOv7 for complex scene SAR images. *Remote Sens.* **2023**, *15*, 2071.
21. Zheng, Z.; Wang, P.; Ren, D.; Liu, W.; Ye, R.; Hu, Q.; Zuo, W. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Trans. Cybern.* **2021**, *52*, 8574-8586.
22. Zhang, H.; Zhang, S. Focaler-iou: More focused intersection over union loss. *arXiv preprint arXiv:2401.10525* **2024**. doi: 10.48550/arXiv.2401.10525.
23. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia, Association for Computing Machinery: New York, NY, USA, 2016; pp. 516-520.
24. Zheng, S.; Luo, X.; Sun, Y. Pedestrian detection algorithm in fog based on improved YOLOv8. *J. Hubei Univ. (Nat. Sci. Ed.)* **2025**, 1-9.
25. Luo, Y.; Yang, W.; Wu, S.; Xu, Z.; Pan, N. Ship target detection algorithm in SAR images based on YOLOv8-CKS. *J. Hubei Univ. (Nat. Sci. Ed.)* **2025**, 1-8.
26. Xiao, Z.; Yan, S.; Qu, H. Safety helmet detection method in complex environment based on multiple mechanism optimization of YOLOv8. *Comput. Eng. Appl.* **2024**, *60*, 172-182.
27. Sun, Y.; Zheng, S.; Luo, X. Sign language recognition algorithm based on improved YOLOv8n. *J. Hubei Univ. (Nat. Sci. Ed.)* **2025**, 1-12.
28. Liu, Z.; Zhang, P.; Guan, X.; Yu, S.; Zhang, X. Wood defect detection and classification based on improved YOLOv8. *For. Eng.* **2025**, 1-16.