MINOR SURFACE DEFECT DETECTION IN AGRICULTURAL MACHINERY USING AN OPTIMIZED YOLOv11N ARCHITECTURE
/
基于改进 YOLOv11n 的农业机械微小表面缺陷检测方法

Min LI, Shuai YUAN*, Wenhong TANG
College of Mechanical and Electronic Engineering, Kunming University of Science and Technology, Kunming 650500/China
Tel: +86-17787109046; E-mail: 2369454071@qq.com
Corresponding author: Shuai Yuan
DOI: https://doi.org/10.35633/inmateh-78-10

Keywords: *agricultural machinery, surface defect detection, YOLOv11n, feature fusion*

ABSTRACT
Reliable operation of agricultural machinery depends on the structural integrity of its steel components during manufacturing. To accurately detect minor surface defects on steel components of agricultural machinery in complex environments, this study proposes an improved detection model, CGC-YOLOv11n, based on the YOLOv11n architecture. The research focuses on enhancing real-time, high-precision defect detection for agricultural machinery maintenance, addressing challenges such as subtle defects under dust, vibration, and foreign object conditions. First, Converse2D reverse convolution is integrated into the C3k2 module to enhance fine-grained feature representation for subtle and blurred defects. Second, the GESA module replaces C2PSA, leveraging dynamic sparse attention to strengthen multi-scale aggregation and focus on key target cues. Third, a Coordinated Detail-Preserving Contextual Fusion (CDPCF) module—an innovative extension of DPCF—is embedded in the neck, employing adaptive content-aware gating to synergistically balance high- and low-resolution features. Experimental results demonstrate that CGC-YOLOv11n achieves a precision of 74.85%, a recall of 74.16%, and a mean average precision (mAP@0.5) of 79.81%, representing an improvement of 1.83 percentage points over the baseline YOLOv11n. Across datasets including NEU-DET and a self-collected farm machinery set, CGC-YOLOv11n delivers superior detection performance. The improved model exhibits robust capabilities in agricultural machinery maintenance, providing technical support for reliable agricultural operations.

摘 要
农机设备的可靠运行取决于其钢制部件在生产过程中的结构完整性.为在复杂环境中精确检测农机钢制部件的微小表面缺陷，本研究基于 YOLOv11n 架构提出改进检测模型 CGC-YOLOv11n。该研究聚焦于提升农机维护中的实时高精度缺陷检测能力，解决尘土、振动和异物等实际场景挑战。首先，Converse2D 反卷积集成到 C3k2 模块中，以增强对细微模糊缺陷的精细特征表征能力。其次 GESA 模块替代 C2PSA，通过动态稀疏注意力机制强化多尺度聚合和对关键目标线索的关注能力，精准聚焦关键目标特征。第三，在颈部嵌入创新性扩展模块 CDPCF（协调细节保留上下文融合），该模块基于 DPCF 模块，采用自适应内容感知门控机制协同平衡高/低分辨率特征。实验结果表明，CGC-YOLOv11n 模型实现 74.85%的精确率、74.16%的召回率及 79.81%的平均精确率（mAP@0.5），较基线 YOLOv11n 提升 1.83 个百分点。在 NEU-DET 及自采集农机数据集测试中，CGC-YOLOv11n 展现出更精准的检测效果。该优化模型在农机维护领域表现出强健性能，为可靠农业作业提供了技术保障。

INTRODUCTION
In recent decades, the convergence of advanced computing and artificial intelligence has revitalized traditional manufacturing. Steel remains a foundational raw material in agricultural machinery, where its quality directly dictates equipment durability, operational reliability, and overall agricultural productivity. Surface defects in steel compromise mechanical integrity, accelerated corrosion, product scrap, and recalls *(Wang et al., 2021*). Thus, accurate and efficient detection of surface defects in agricultural machinery is essential to prevent field failures, lower maintenance costs, and support sustainable crop production.

Defect detection techniques fall into two main categories: traditional and deep learning paradigms associated with computer vision. Traditional methods primarily rely on manual observation, which is time-consuming, labor-intensive, and often ineffective due to the subtle nature or small defects that are hard to discern visually. With the widespread adoption of computers, deep learning has gained prominence in defect detection (*Zhipeng et al., 2025*). Convolutional neural network (CNN) has attracted sustained scholarly interest, outperforming traditional methods across various visual tasks (*Li et al., 2024*). In quality inspection, deep learning object detection has evolved from two-stage methods, such as Faster R-CNN (*Ren et al., 2015*), to single-stage methods, including RetinaNet (*Lin et al., 2017*) and the YOLO series (*Xu et al., 2024*). The former offer high accuracy and precise localization, making them suitable for complex scenes, but they are computationally intensive and slow. The latter, by contrast, are fast and structurally simple, ideal for real-time applications, though they typically exhibit lower accuracy, particularly for small or densely packed targets.

Building on the above foundations, scholars have proposed improvements to object detection frameworks. For instance, *Diogo et al. (2025)* proposed a low-cost modular multimodal solution for PCB solder joint inspection, integrating precise scanning with an electric XY platform, synchronized RGB/thermal imaging, and deep learning. They incorporated cross-scale attention enhancements like FSPPCSP and HCAM into YOLOv11n to detect hidden defects like cold solder joints. *Shi et al., (2025)*, proposed an improved YOLOv9s-based wafer defect detection algorithm, incorporating inverted channel attention mechanism, SPD-conv downsampling, dynamic head, and focal loss to handle unbalanced datasets and diverse defect patterns. Meanwhile, efficient surface defect detection is crucial for the structural health monitoring and predictive maintenance of agricultural machinery (*Bai et al., 2025*). Like general industrial steel, components in farming equipment are often assembled in industrial environments, necessitating robust diagnostic tools to ensure operational reliability. *Tian et al. (2020)* summarizes the latest applications of computer vision in agricultural automation—including crop monitoring, pest and disease control, automated harvesting, quality inspection, farm management, maintenance of agricultural machinery production, and UAV monitoring. It analyzes challenges and explores how integrating deep learning will advance the development of smart agriculture.

Although both two-stage and single-stage algorithms have made substantial progress in detecting various objects, in agricultural production, steel defect in machinery components can cause operational failures, leading to downtime and yield losses. While deep learning has advanced defect detection in industrial settings, its application to agricultural machinery quality control remains limited. This work adapts YOLOv11 *(Khanam & Hussain, 2024*) for minor defect detection in steel parts used in farming equipment. Based on these considerations, this paper proposes an improved small-target detection method, the CGC-YOLOv11n model, built on the latest YOLOv11n framework to enhance overall performance.

MATERIALS AND METHODS

Dataset Preparation

The images for this study were sourced from the NEU-DET (*He et al., 2020*) and supplemented with images collected from an agricultural machinery assembly plant. These defects share identical physical and morphological characteristics with structural deterioration observed on steel components of agricultural machinery operating under high vibration, soil abrasion, cyclic loading, and humid environments. Minor surface defects such as patches, inclusions, scratches, crazing, pitted surface, and rolled-in scale may progressively propagate under field operating conditions, leading to fatigue failure or structural weakening. Therefore, early detection of minute surface defects is critical for ensuring the long-term operational reliability and safety of agricultural machinery systems. To specifically address the quality control requirements for agricultural machinery, images were carefully selected to simulate common defect patterns encountered during production and subsequent maintenance, totaling 1,873 images. After removing images that could not be classified as a specific defect type, the remaining unlabeled images were manually annotated using the LabelMe tool, providing precise bounding boxes and category labels for each defect. This comprehensive preparation ensures the model's capacity to localize minor structural defects under the unpredictable conditions characteristic of agricultural manufacturing and maintenance. This resulted in a final dataset of 1,800 images, which was split into a training set of 1,260 images, a validation set of 270 images and a test set of 270 images in a 14:3:3 ratio. Targeted data augmentation, including Gaussian noise and brightness adjustment, was applied to simulate environmental interference, such as soil contamination, rust, and varying outdoor illumination, characteristic of orchard and field maintenance scenarios.

Detection Model Based on Improved YOLOv11n

The experiments in this study were conducted on a Windows 10 operating system with an Intel Core i5-12600KF CPU (3.70 GHz) and 32 GB of RAM. The graphics processing unit (GPU) was an NVIDIA GeForce RTX 4060 Ti with 16 GB of video memory. The programming language was Python 3.10, and the PyTorch 2.3 deep learning framework was used for model construction and training. CUDA 12.1 was used to accelerate the deep learning algorithms. The AdamW optimizer was adopted. The batch size was 32, the number of epochs was 300, and the number of workers was 4. The initial learning rate was 0.003, and the input image size was 640 × 640 × 3. Other settings, including data augmentation, used the default configurations of Ultralytics.

YOLOv11 continues the end-to-end design of the YOLO series, including four parts: Input layer, Backbone, Neck, and Head. Among them, the lightweight version YOLOv11n reduces the network depth and width to lower the computational complexity and improve the inference efficiency. The defect detection model CGC-YOLOv11n proposed in this study optimizes the structure of the original YOLOv11n model. The improved model CGC-YOLOv11n performs better than the previous model in small object detection and simultaneously achieves a balance between model lightness and performance.

Firstly, C3k2-Converse is constructed based on the C3k2, which is used to replace some C3k2 modules in the backbone and neck network to improve the ability to capture the key information of the input. Then, at the backbone output, the Gated Elastic Sparse Aggregation (GESA) module, based on the original C2PSA architecture, effectively enhances the aggregation ability of multi-scale features by optimizing and improving the self-attention mechanism through the Efficient Prompt Guide Operator (EPGO), thereby improving the model's perception and localization capabilities for targets in complex scenarios. Last, the Coordinated Detail-Preserving Contextual Fusion (CDPCF) to replace Concat, which dynamically integrates multi-scale features via learnable fusion strategies, employing a gating mechanism to optimally weight the contributions from both high- and low-resolution features. The improved network structure diagram is illustrated in Figure 1.

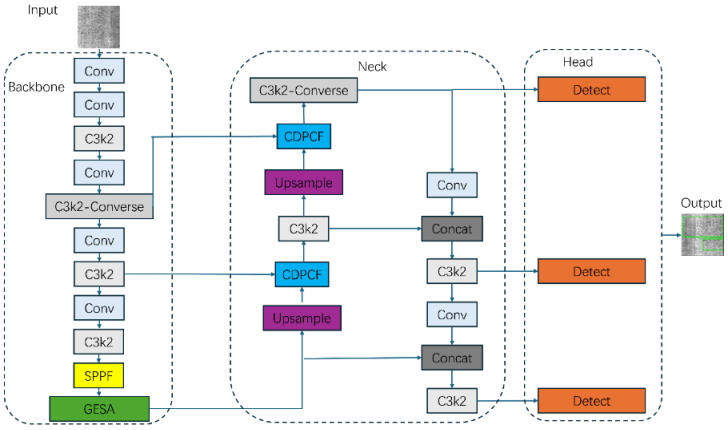


Fig. 1 – CGC-YOLOv11 network architecture diagram

C3k2-Converse

During the assembly processing of agricultural equipment, long, slender cracks can form in a meandering and branching pattern owing to stress concentration or material inhomogeneity. Additionally, when steel equipment is scratched by foreign objects, it may exhibit low-contrast curved line structures. The C3k2 module uses Conv, which is usually accompanied by downsampling in the deeper layers of the network, leading to a reduced spatial resolution and loss of fine-grained features. To tackle this issue, this paper introduces the Converse2D *(Huang et al., 2025)* module to improve the C3k2 module. The standard C3k2 module is replaced with C3k2-Converse(kernel=3x3) at the 4th layer of the backbone and the 16th layer of the neck. The improved C3k2-Converse is depicted in Figure 2.

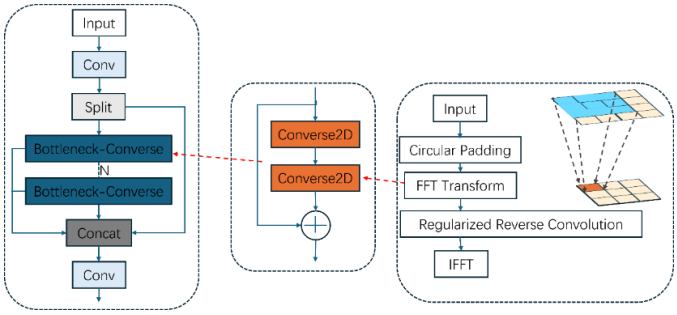


Fig. 2 – C3k2-Converse structure diagram

Converse2D is a deep reverse convolution operator that reconstructs the input feature map X from the observed output Y, given the kernel K and scale factor s. This can help with deblurring and super-resolution, enabling the C3k2-Converse module to better detect certain blurry and tiny defects, as it explicitly models the inverse process in the feature space.

$$X^* = \arg\min_X \left\| Y - (X \otimes K) \downarrow_s \right\|_F^2 \tag{1}$$

where X is the input feature map that needs to be reconstructed; Y is the observed output, obtained after convolution⊗and downsampling↓s; $\|.\|F$ represents the Frobenius norm. Since the zero values of kernel K at certain frequencies can cause the solution to be infinitely amplified in the frequency domain, leading to numerical divergence or instability, a quadratic regularization term is introduced to stabilize the results.

$$X^* = \arg\min_X \left\| Y - (X \otimes K) \downarrow_s \right\|_F^2 + \lambda \left\| X - X_0 \right\|_F^2 \tag{2}$$

where λ is a regularization parameter that ensures mathematical stability by penalizing excessive amplification of noise in high-frequency components, preventing unstable solutions in this ill-posed inverse problem, avoiding overfitting to noise and promoting solutions close to the initial estimate X_0. To solve the optimization problem in Eq.(2), the computation is transferred to the frequency domain, a closed-form solution is obtained in the frequency domain under cyclic boundary conditions (*Zhao et al., 2016*), resulting in scaling for s > 1 and no scaling for s = 1.

$$X^* = F^{-1}\left(\frac{1}{\lambda} \left(L - \overline{F_K} \odot_s \frac{(F_K L) \Downarrow_s}{|F_K|^2 \Downarrow_s + \lambda} \right) \right) \tag{3}$$

$$L = \overline{F_K} F_{Y \uparrow_s} + \lambda F_{X_0} \tag{4}$$

$$X^* = F^{-1}\left(\frac{\overline{F_K} F_Y + \lambda F_{X_0}}{|F_K|^2 + \lambda} \right) \tag{5}$$

where $F(\cdot)$ denotes the fast Fourier transform (FFT), and $F^{-1}(\cdot)$ represents the inverse one. $IF_kI^2 = F_k \odot F_k$ denotes the elementwise squared magnitude. If IF_kI^2close to 0, without λ, the denominator will be very small, leading to the solution will diverge; $\odot_s$ multiplication of different blocks in s × s. The input x is upsampled, and it is determined whether to scale x. The FFT is computed, the inverse is normalized, and the inverse FFT is finally performed. This is contrary to conventional transposed convolution.

$$Y = (X \otimes K) \downarrow_S \tag{6}$$

where Converse2D solves an ill-posed inverse problem in frequency domain (FFT-based closed-form), recovering high-frequency details lost in downsampling-crucial for blurred cracks- while λ regularization ensures stability. Unlike transposed conv, it explicitly reconstructs features, reducing artifacts in fine defects.

GESA

When performing object detection tasks, the C2PSA module calculates the correlation score of each point in the image with all other points. Although this method can capture global relationships, it can lead to computational redundancy, information noise, and other issues. Inspired by the role of the EPGO module in the CPRA former network structure (*Zou et al., 2025*), C2PSA was improved, and GESA was designed. As shown in Figure 3, the proposed GESA module introduces an attention-guided mechanism that allows the module to focus only on the most important cues, effectively avoiding the issues exposed by the C2PSA module, thereby improving the model performance.

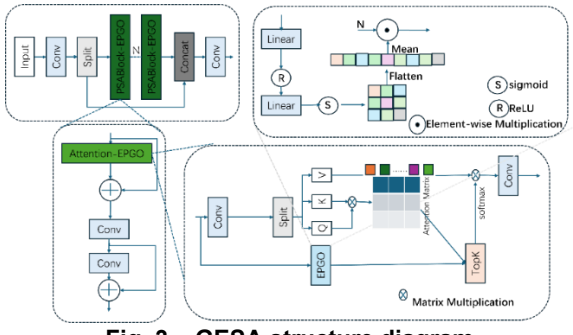


Fig. 3 – GESA structure diagram

The input feature map $X \in R^{C\times H\times W}$ is linearly transformed through a 1x1 convolution layer to generate query(Q), key(K), and value(V). Then, Q and K perform matrix multiplication to obtain the attention matrix. Meanwhile, the input image generates a dynamic K $\in$(0, N) value through $EPGO$, K is dynamically predicted per image. Unlike sparse attention with a fixed $TopK$ value *(Jiang et al., 2025*) or a fixed sparsity pattern, $EPGO$ dynamically modulates sparsity levels in accordance with the input image content (e.g., areas with complex textures or smoothness). For regions rich in detail, more tokens may be retained (larger K value), and for areas with simple backgrounds, only a few key tokens may be focused on (smaller K value), enabling adaptive multi-scale perception.

$$[M_k]_{ij} = \begin{cases} 1, & \text{if } j \in \Omega_i^{(k)} \\ 0, & \text{otherwise.} \end{cases} \Leftrightarrow M_{ij} \in \mathrm{Top}_k(M_{i,:}), \tag{7}$$

where $\Omega*$ represents the k most important elements of the i-th row of the attention matrix. By sparsifying, this module reduces the interference of invalid or noisy tokens in the attention calculation, enabling the model to concentrate on the most important feature relationships, outperforming fixed $Topk$ by focusing on subtle anomalies. Simultaneously, it avoids the need to compute the full N×N attention matrix later, theoretically saving computational resources. This improvement enhances the feature extraction capability. Finally, the attention matrix processed using Eq.(7) through $Softmax$ and performs matrix multiplication with the value V.

$$Attention - EPGO = \mathrm{Softmax}(\mathrm{Topk}(\frac{QK^T}{\sqrt{d_k}}))V \tag{8}$$

where $Topk$ represents the selection operation on the attention matrix after $EPGO$; d_k is a weight coefficient, the inner product of vectors Q and K, used to obtain the relationship between Q and k, and then the final result is obtained through convolution and then a projection.

CDPCF

Since the defect is small target, some small defect target information may be diluted or lost during neck feature fusion. Unlike the original DPCF *(Xu et al., 2025)*, which employs learned parameters to generate weights, the improved module in this paper adopts Adaptive Content-Aware Gating Fusion. By connecting high- and low-resolution features and leveraging 1×1 convolutions to extract cross-scale correlations, the model dynamically achieves an optimal balance between detail preservation and contextual integration based on local features. As shown in Figure 4, CDPCF replaces Concat at layers 12 and 15, where high-resolution shallow features meet low-resolution deep semantics—critical for preventing small-defect dilution in final P3/P4 heads. It employs a learnable fusion strategy and utilizes a spatial adaptive gating mechanism to balance the contributions of high- and low-resolution features, thereby enhancing detail preservation and contextual understanding ability in small target detection.

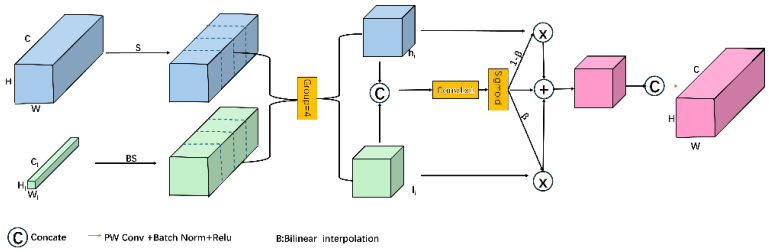


Fig. 4 – CDPCF structure diagram

Firstly, it is necessary to align high- F_h and low-resolution F_l features in the spatial dimension. Then, split the two channel-aligned feature maps into four equal segments each, with each segment of the two feature maps belonging to $R^{HxWxC/4}$. Concatenate the high-resolution feature component h_i with the low-resolution feature component l_i, then pass it through a $1x1$convolutional layer, and finally apply the Sigmoid activation function. This design incorporates the concept of an attention mechanism. The weight $\beta \in [0,1]$ is now dynamically computed based on the content of the two feature layers. This enables the model to adjust the fusion ratio in real time depending on whether the current region belongs to a “detail-rich object” or a “complex background.”

$$\beta = sigmoid(Conv_{1\times1}([h_i; l_i])), \quad \beta \in R^{H\times W\times\frac{C}{4}} \tag{9}$$

The contributions of all segments are integrated through a weighted sum for each segment

$$o_i^{'} = \beta \odot l_i + (1-\beta) h_i \tag{10}$$

where ⊙ denotes element-wise multiplication, and β is used to control, balance context and details adaptively. This weighted fusion process ensures that the contributions of the low- and high-resolution features are adaptive during fusion, thereby preventing information loss. Each segment o* is integrated to obtain $F^*\in R^{HxWxC}$. Finally, F* is processed through convolution, batch normalization, and $RELU\ (\delta\ (.))$ to obtain the output F_o of this module, further integrate features and suppress noise.

$$F_O^{'} = \left[o_1^{'}, o_2^{'}, o_3^{'}, o_4^{'}\right] \tag{11}$$

$$F_O = \delta(B(Conv(F_O^{'}))) \tag{12}$$

RESULTS

Experimental Design and Analysis

Evaluation Metrics

This paper used standard metrics to evaluate performance, including Parameters, Precision, Recall, F1 Score, mean Average Precision (mAP), model parameters, and GFLOPs to assess the model performance (*Lyu et al., 2025*). These results demonstrate the model's capacity to learn defect patterns, convergence behavior, and potential limitations related to the data scale.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{13}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{14}$$

$$F1\ Score = \frac{2*\text{Precision}*\text{Recall}}{(\text{Precision}+\text{Recall})} \tag{15}$$

where: TP is the number of true positives; FP is the number of false positives; FN is the number of false negatives.

The P-R curve area enclosed by the curve and the axis represents the average precision AP of the defects in this category. By calculating the AP of the defects in all categories and averaging them, the mean AP average precision value can be obtained. Among them, mAP@0.5 and mAP@0.5:0.95 are used to comprehensively evaluate the model performance at different IoU thresholds. Where N is the total number of classes, and AP_i is the corresponding defect class.

$$AP = \int_0^1 P(R)dR \tag{16}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{17}$$

Ablation Experiment

This paper examines the detection accuracy and model size changes of the improved CGC-YOLOv11n, in the self-built dataset, and designs 7 ablation experiments to verify its effectiveness in the task of defect recognition. The results are presented in Table 1.

Table 1

Results of model ablation experiments

Model	Parameters/10^6	GFLOPs	Recall/%	F1 Score/%	Precision/%	mAP@0.5/%
YOLOv11n	2.58	6.3	73.31	71.97	71.56	77.98
YOLOv11n+C3k2-Trans Conv	2.57	6.2	74.32	73.09	74.41	78.79
YOLOv11n +C3k2-Converse	2.57	6.1	75.02	75.21	76.39·	79.07
YOLOv11n +CDPCF	2.60	6.3	74.58	74.76	75.71	78.45

Model	Parameters/10^6	GFLOPs	Recall/%	F1 Score/%	Precision/%	mAP@0.5/%
YOLOv11n +GESA	2.59	6.3	76.81	74.82	73.28	78.56
YOLOv11n+C3k2-Converse+CDPCF	2.58	6.1	74.46	73.98	74.7	79.52
YOLOv11n+C3k2-Converse +CDPCF+GESA	2.59	6.1	74.16	74.19	74.85	79.81

As shown in Table 1, each improvement method enhances the model's detection performance to some extent. Compared with YOLOv11n, the use of C3k2-Converse improved all performance indicators, particularly precision, recall, F1-score, and mAP@0.5, while reducing the computational cost by 0.2 GFLOPs. This indicates that detail reconstruction performs better than transposed convolution, and that the addition of extra components enhances the object detection capability of the model. And even compared to transposed convolution, the effect of Converse2D is better. After introducing CDPCF, the model's mAP@0.5 is 78.45%, and the optimization effect of the module is not as good as that of C3k2-Converse, but it still shows an improvement over the original model. When the backbone is constructed with GESA, the model's mAP@0.5 is 78.56%, which is 0.58% higher than the original YOLOv11n model, suggesting that GESA enhances feature extraction and improves the overall quality of defect features. Combining C3k2-Converse and CDPCF, the model's mAP@0.5 is 79.52%; however, there remains potential for further optimization. The subsequent integration of the GESA module leading to the higher performance. This suggests that GESA's dynamic sparse attention mechanism effectively filtered redundant features generated by the other 2 modules, enabling a more synergistic fusion. After integrating all the above improvements, the mAP@0.5 of CGC-YOLOv11n is 79.81%, with no significant changes in the number of parameters and GFLOPs, and the mAP@0.5 has increased by 1.83% compared with the base model, proving that the CGC-YOLOv11n model has good accuracy and applicability in defect recognition tasks.

To better understand the contribution of the model's effective receptive field (*Ding et al., 2022*), the validation set was used to test the ratio of the model's output final feature center points and the corresponding regions in the input images. The feature extraction outputs from the backbone to the neck were selected to visualize the size of the model's receptive field. As illustrated in Figure 5, darker colors indicate higher correlation. Without changing the network structure, the improved model CGC-YOLOv11n achieves a larger effective receptive field. A larger receptive field can better capture contextual information, see more parts of the detected object, and provide stronger robustness with greater resistance to local interference, because, unlike the baseline model, its receptive field is not concentrated and does not depend on a single local region.

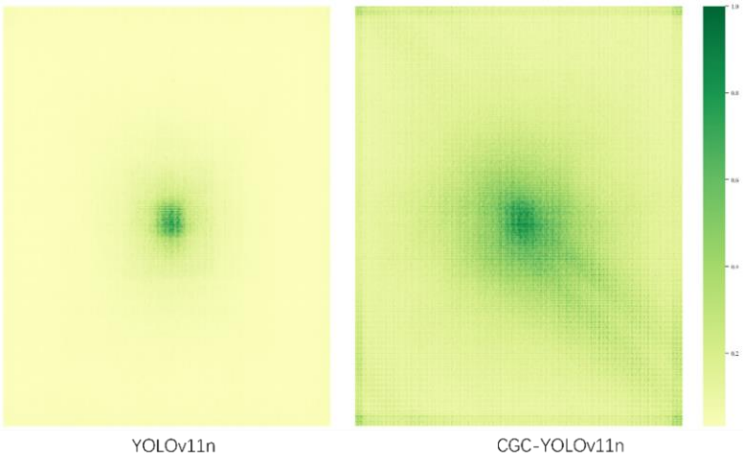


Fig. 5 – Visualization of effective receptive field size for each model

Comparison Experiments

To further assess model performance, it was compared with mainstream models, including Faster R-CNN, YOLOv5s, SSD (Liu et al., 2015), RetinaNet, YOLOv6n, FCOS (Tian et al., 2019), YOLOv8n, YOLOv10n, and YOLOv12n, on the same dataset. The experimental results are shown in Table 2. All models were trained from scratch.

Table 2

Performance comparison of mainstream models

Model	Precision/ %	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	Parameters/10^6	FPS
FasterR-CNN	73	75.5	78.2	42.7	41.37	73.8
YOLOv5s	60.8	71.7	68.5	31	7.03	136
SSD	73	70.1	74.6	38.2	24.41	91
RetinaNet	73.7	75	78.6	40.2	36.43	99.8
YOLOv6n	73.2	70.21	76.63	42.84	4.23	271.31
FCOS	68.9	75.3	76	38.1	32.13	90.8
YOLOv8n	75.94	74	76.64	43.53	3.01	208.09
YOLOv10n	73.43	74.47	76.79	42.88	2.27	254.97
YOLOv12n	66.72	76.02	75.97	44.15	2.56	211.08
CGC-YOLOv11n	74.85	74.16	79.81	45.45	2.59	231.65

As shown in Table 2, CGC-YOLOv11n outperforms mainstream methods on the dataset. The two key metrics, mAP@0.5 and mAP@0.5:0.95, achieved the best results, reaching 79.81% and 45.45%, respectively. At the same time, the model has only 2.59 M parameters, which is slightly higher than YOLOv10n and YOLOv12n, but it still maintains a lightweight design, indicating that the introduced improvement modules do not significantly increase the model complexity. As a classic two-stage model, Faster R-CNN achieves an mAP@0.5 of 78.2%, but its number of parameters reaches 41.37 M, which limits its applicability in real-time detection scenarios. Meanwhile, CGC-YOLOv11n achieves an mAP@0.5 of 79.81%, representing improvements of 1.61%, 11.31%, 5.21%, 1.21%, 3.18%, 3.81%, 3.17%, 3.02%, and 3.84% compared with Faster R-CNN, YOLOv5s, SSD, RetinaNet, YOLOv6n, FCOS, YOLOv8n, YOLOv10n, and YOLOv12n, respectively. Although its FPS is lower than that of several YOLO variants, it still meets real-time requirements and provides a favorable trade-off between accuracy and computational complexity. Overall, its performance surpasses that of other mainstream object detection algorithms, demonstrating strong detection capability.

After the ablation and comparison experiments, additional experiments were conducted on the GC10-DET dataset *(Lv et al., 2020)*, with the same data distribution as the self-built dataset, to verify the generalization ability of the proposed network model. The CGC-YOLOv11n model was directly transferred to this dataset, and a comparative experiment with YOLOv11n was conducted. The experimental results are shown in Table 3. Compared with YOLOv11n of the same scale, the proposed model achieved better overall detection accuracy under similar computational cost and parameter size, with an improvement of 0.98% in mAP@0.5, validating the effectiveness and robustness of the proposed method.

Table 3

Experimental results on GC10-DET

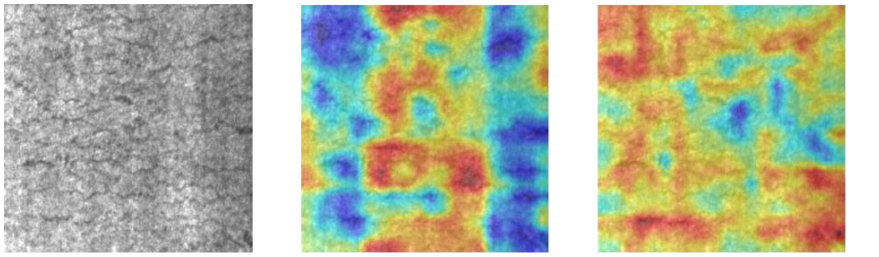
Model	Parameters/ 10^6	GFLOPs	Recall/ %	F1 Score / %	Precision / %	mAP@0.5/%	mAP@0.5: 0.95/%
YOLOv11n	2.59	6.1	63.2	63.17	63.68	63.67	31.65
CGC-YOLOv11n	2.58	6.3	60.18	63.9	71.3	64.65	33.07

To evaluate the practical application of CGC-YOLOv11n in agricultural machinery, the model was deployed on the RK3588 NPU edge device (Figure 6), achieving 81 FPS under 640x640 input resolution. This enables real-time detection on assembly lines for farm equipment. The imaging setup included a high-resolution camera, with an acquisition distance of 0.5–1 m. For maintenance diagnostics in simulation experiments, the model was tested on handheld scans of field-exposed components images, results showed 12.35 ms inference time. These results confirm the model's potential for automated inspection stations in factory, reducing downtime.

Fig. 6 – RK3588

Visualization of Model Recognition Effect

To more intuitively observe the characteristics of the attention regions when the CGC-YOLOv11n model recognizes surface defects and verify the effectiveness of the improved model, the Grad-CAM (Gradient-weighted Class Activation Mapping) *(Selvaraju et al., 2016)* method was adopted to visualize the recognition effects of the YOLOv11n and CGC-YOLOv11n models (Figure 7).



a. Original image b. YOLOv11n heat map c. CGC-YOLOv11n heat map

Fig. 7 – Comparison of heat maps between two models

In Fig. 7b, the heatmap's areas of focus are more scattered, where the heatmap does not accurately concentrate on the edges or key areas of the steel defects, indicating that the model has weak attention to the defect areas, and there may be instances of missed or false detections. However, in Fig. 7c, the areas of focus in the heatmap are more concentrated, with the defect areas being brighter and more focused, showing that the improved model can better identify and concentrate on the defect areas, demonstrating higher detection accuracy and attention. This suggests that the improved model CGC-YOLOv11n performs better in detecting defects, being able to locate and identify defect areas more accurately. Figure 8 shows the visualization results of the proposed algorithm for defect target detection presented in this paper.

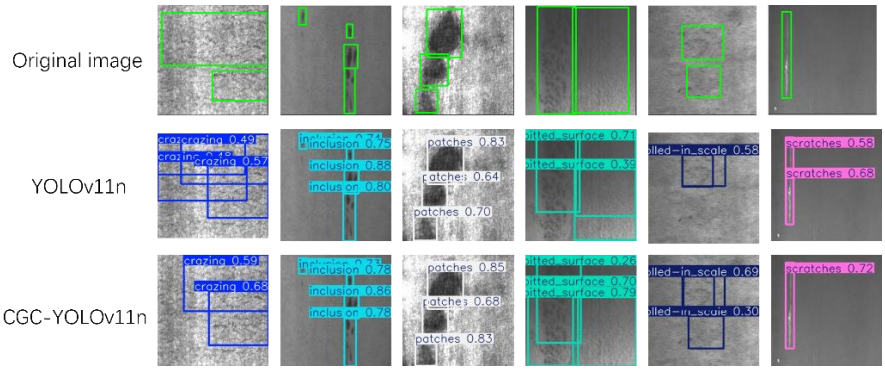


Fig. 8 – Comparison of proposed algorithm and original model in visualization results

From the detection results, it can be seen that the algorithm proposed in this paper detected some targets that the original YOLOv11n framework failed to detect. This indicates that, compared to YOLOv11n, the algorithm in this paper enhances the detection capability for small and dark-toned defective targets through the improved C3k2 module and CDPCF module, which to some extent alleviates interference from complex backgrounds and also reduces false detections. In summary, the CGC-YOLOv11n algorithm proposed in this paper has achieved good detection results in defect detection scenarios.

CONCLUSIONS

This study focuses on the problem of minor surface defect detection in agricultural machinery components, where defects are typically characterized by small size, weak texture, and low contrast. CGC-YOLOv11n based on YOLOv11 detection method is developed by improving feature representation and multi-scale feature fusion, enabling more effective preservation of defect details under complex backgrounds. The core idea is 'preserve details—attention mechanism—enhanced fusion': structurally, the convolutions in the residual blocks of the C3k2 module are replaced with Converse2D, combined with the feature fusion of the GESA module guided by EPGO features; the CDPCF module is used to improve the network's neck, enhancing the overall performance of the model. Experimental results show that, compared with the baseline, the improved algorithm achieved increases of 1.83% in mAP@0.5 on the self-built dataset, with a Precision improvement of 3.29%. Its number of parameters 2.59 M and GFLOPs 6.1 remain within the range of lightweight models. Compared with current mainstream algorithms, CGC-YOLOv11n achieves the balance between detection accuracy and inference speed.

In summary, CGC-YOLOv11n systematically enhances detection capability and localization accuracy for small defects in surface detect without significantly increasing computational overhead. It can reduce manual inspection errors and minimize losses, addressing key pain points in agricultural machinery manufacturing and maintenance. Future research will explore more efficient feature fusion strategies and aiming to meet high-precision detection requirements.

REFERENCES

[1] Bai, X., Chen, Q., Song, X., & Hong, W. (2025). Advancing agricultural machinery maintenance: Deep learning-enabled motor fault diagnosis. *IEEE Access*.

[2] Ding, X., Zhang, X., Zhou, Y., Han, J., Ding, G., & Sun, J. (2022). Scaling Up Your Kernels to 31×31: Revisiting Large Kernel Design in CNNs. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11953-11965.

[3] Diogo, T., Ramos, A., Pereira, F., Araújo, N., & Lopes, A. (2025). Automated detection and classification of soldering defects in printed circuit boards using deep learning and optical and thermal imaging. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-025-02691-5

[4] He, Y., Song, K., Meng, Q., & Yan, Y. (2020). An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE transactions on instrumentation and measurement*, *69*(4), 1493-1504.

[5] Huang, X., Liu, S., Zhang, K., Tai, Y., Yang, J., Zeng, H., & Zhang, L. (2025). Reverse Convolution and Its Applications to Image Restoration. Proceedings of the IEEE/CVF International Conference on Computer Vision,

[6] Jiang, Y., Wu, Y., & Zhao, B. (2025). Enhancing SLAM algorithm with Top-K optimization and semantic descriptors. *Scientific Reports*, *15*(1), 8280. https://doi.org/10.1038/s41598-025-90968-3

[7] Khanam, R., & Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

[8] Li, C., Zhang, Y., Shi, Z., Zhang, Y., & Zhang, Y. (2024). Moderately dense adaptive feature fusion network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1-12.

[9] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision,

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C.-Y., & Berg, A. C. (2015). SSD: Single Shot MultiBox Detector. European Conference on Computer Vision,

[11] Lv, X., Duan, F., Jiang, J.-j., Fu, X., & Gan, L. (2020). Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network. *Sensors*, *20*(6), 1562. https://www.mdpi.com/1424-8220/20/6/1562

[12] Lyu, D., Li, X., Wang, W., Wu, B., Shi, S., & Shen, H. (2025). A melon fruit diameter measurement method based on an improved mask R-CNN. *INMATEH-Agricultural Engineering*, *77*(3), pp.115-125. DOI: https://doi.org/10.35633/inmateh-77-09.

[13] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

[14] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*, 336 - 359.

[15] Shi, T., Wang, X., & Mao, J. (2025). A Wafer Defect Detection Method for Unbalanced Data. *Journal of Failure Analysis and Prevention*, *25*(4), 1694-1705. https://doi.org/10.1007/s11668-025-02227-2
[16] Tian, H., Wang, T., Liu, Y., Qiao, X., & Li, Y. (2020). Computer vision technology in agricultural automation—A review. *Information processing in agriculture*, *7*(1), 1-19.
[17] Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully Convolutional One-Stage Object Detection. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9626-9635.
[18] Wang, Y., Guo, Z., Bai, X., & Yuan, C. (2021). Effect of weld defects on the mechanical properties of stainless-steel weldments on large cruise ship. *Ocean Engineering*, *235*, 109385.
[19] Xu, W., Zheng, S., Wang, C., Zhang, Z., Ren, C., Xu, R., & Xu, S. (2025). SAMamba: Adaptive state space modeling with hierarchical vision for infrared small target detection. *Information Fusion*, *124*, 103338.
[20] Xu, Y., Li, J., Dong, Y., & Zhang, X. (2024). Survey of development of YOLO object detection algorithms. *J. Front. Comput. Sci. Technol*, *18*(09), 2221-2238.
[21] Zhao, N., Wei, Q., Basarab, A., Dobigeon, N., Kouamé, D., & Tourneret, J.-Y. (2016). Fast Single Image Super-Resolution Using a New Analytical Solution for ℓ2 – ℓ2 Problems. *IEEE Transactions on Image Processing*, *25*(8), 3683-3697.
[22] Zhipeng, D., Shiming, H., Gen, Y., & Junfen, M. (2025). Survey of texture surface defect detection method based on deep learning. *Computer Integrated Manufacturing System*, *31*(3), 721-745.
[23] Zou, S., Zou, Y., Li, J., Gao, G., & Qi, G. (2025). Cross Paradigm Representation and Alignment Transformer for Image Deraining. *arXiv preprint arXiv:2504.16455*.