Automation and Machine Learning (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2025.060119
ISSN 2516-5003 Vol. 6 Num. 1

Metal Defect Detection Algorithm Based on Improved YOLOv11

Ruiming Liu[1,a], Yunliang Du[1,b,*], Xuesong Duan[1,c], Shuai Huang[1,d], Yong Liu[1,e], Zhifei Wang[2,f]

[1]School of Electronic Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu, China
[2]School of Mechanical and Electrical Engineering, Lianyungang Technical College, Lianyungang, Jiangsu, China
[a]liurm@jou.edu.cn, [b]2023220622@ jou.edu.cn, [c]2023210601@ jou.edu.cn,
[d]2023220626@jou.edu.cn, [e]897771817@qq.com, [f]19826123239@163.com
[*]Corresponding author

Keywords: YOLOv11; Defect Detection; Convolution; Feature Fusion

Abstract: In industrial manufacturing, the inspection of metal products holds significant importance due to the detrimental impact defects such as corrosion, welding issues, holes, cracks, among others, can have on the functionality and longevity of metal components. Conventional methods for detecting metal defects suffer from drawbacks including low efficiency, heavy reliance on human intervention, and limited adaptability to complex environments, thereby falling short of the requirements for modern high-precision and automated detection. To address these challenges, this study introduces an enhanced YOLOv11-AAG model, building upon the YOLOv11 framework, aimed at enhancing the precision and effectiveness of metal defect identification. The enhancements to the original YOLOv11 architecture primarily focus on three key areas: feature extraction, feature fusion network, and detector design. Comparative analysis with YOLOv8, Faster R-CNN, and the baseline YOLOv11 model reveals that the YOLOv11-AAG model achieves an average accuracy of 80.3%, surpassing the 77.1% accuracy of the YOLOv11 model by 3.2%.

1. Introduction

In industrial manufacturing, ensuring the quality of metal products is essential for their performance and longevity. Defects in metals, both on the surface and internally, such as corrosion, welding imperfections, holes, and cracks, can significantly impact the mechanical properties of metal components. Moreover, in industries like aerospace, rail transit, and high-end equipment manufacturing, these defects can pose serious safety risks[1]As intelligent manufacturing and industrial automation advance rapidly, conventional inspection methods like manual visual checks, ultrasonic testing, and eddy current testing are facing challenges. These methods exhibit drawbacks such as low detection efficiency, reliance on human labor, subjective interpretation of results, and limited adaptability to complex environments. Consequently, they struggle to meet the demands of modern industrial production for high precision, automation, and real-time detection.

In recent years, advancements in computer vision technologies, particularly deep learning, have

significantly enhanced object detection capabilities. The YOLO algorithms[2]have emerged as promising tools for industrial defect detection, offering notable advantages in speed and accuracy. Convolutional neural networks, renowned for their adept feature extraction, have garnered significant attention in defect detection applications. Deep learning approaches for defect detection are typically categorized into single-stage and two-stage algorithms[3]. Two-stage algorithms like Faster R-CNN[4] employ a Region Proposal Network (RPN) to identify candidate regions for subsequent classification and bounding box regression. While these methods achieve high accuracy, the intricate candidate region processing hinders real-time detection speed. On the other hand, single-stage detectors such as YOLO and SSD[5] enable direct end-to-end prediction of object attributes through the feature extraction network, demonstrating notable speed advantages. YOLOv11[6] has emerged as a popular choice for metal defect detection, offering a balanced speed-accuracy trade-off.

DD-YOLO[7] optimizes model complexity through knowledge distillation and differentiable architecture search but employs a fixed ratio design in the anchor box mechanism, hindering its ability to handle diverse defect shapes like corrosion and cracks. Conversely, ESI-YOLOv8 enhances computational efficiency through the EP module and SPPF-LSKA module but struggles to address the challenge of preserving defect edge details in complex lighting conditions.The enhanced YOLOv8 model integrates MobileViTv2 and Transformer architectures to improve feature extraction[8]. Nevertheless, a bottleneck persists in the accuracy of its anchor box generation strategy for irregular defects, leading to an mAP of only 74.1%. Similarly, YOLOv11 faces limitations in this context, particularly when detecting metal defects .

In this study, we introduce a novel model, YOLOv11-AAG, to overcome current methodological constraints and leverage state-of-the-art technologies. We devise the C3k2-AKConv[9] module by integrating an edge detail enhancement module, AKConv, based on Sobel convolution into C3k2. This integration aims to enhance the capture of defect edges and textures. Subsequently, we incorporate an Adaptive Multi-scale Feature Fusion Network[10] into the neck structure. This network employs spatial and channel attention mechanisms to focus on minute defects on steel surfaces and utilizes a dynamic routing mechanism to adaptively merge features from various levels. Lastly, to address the challenge of low anchor box matching efficiency resulting from the diverse geometric shapes of metal defects, we introduce an Adaptive Defect-Aware Anchor Box Generation Mechanism (Guided-Anchoring) in the detection head segment. This mechanism dynamically adjusts anchor box size, ratio, and distribution density to precisely accommodate complex defect shapes. Comparative analysis with the YOLOv11 model demonstrates enhanced detection accuracy, convergence speed, and robustness of the proposed model.

2. Principles of Correlation

2.1 YOLOV11 algorithm

YOLOv11, the most recent iteration of the YOLO series, was officially launched by Ultralytics in 2024. It upholds the efficient detection capabilities synonymous with the YOLO series while introducing significant advancements in network architecture, training methodologies, and deployment optimization. The network architecture can be referenced in Figure 1.

2.2 Problems with algorithms

The selection of the YOLOv11 algorithm model was based on its real-time performance and accuracy. However, challenges have arisen in its practical application[11]. YOLOv11 exhibits limitations in capturing edge and texture features effectively. Specifically, its C3k2 module lacks a

targeted edge enhancement mechanism, leading to a weak capability in extracting defect features with low contrast and fuzzy boundaries. Additionally, the traditional FPN+PAN structure used in the neck network relies on fixed weights for feature fusion[12], making it challenging to adaptively focus on local details of small defects on metal surfaces. Furthermore, the anchor frame generation mechanism demonstrates poor adaptability to defect morphology. The fixed density, size, and proportion of anchor frames cannot be dynamically adjusted based on the actual distribution of defects during metal defect detection[13]. Consequently, this results in low matching efficiency between anchor frames and metal defect boundaries, thereby impacting positioning and classification accuracy.

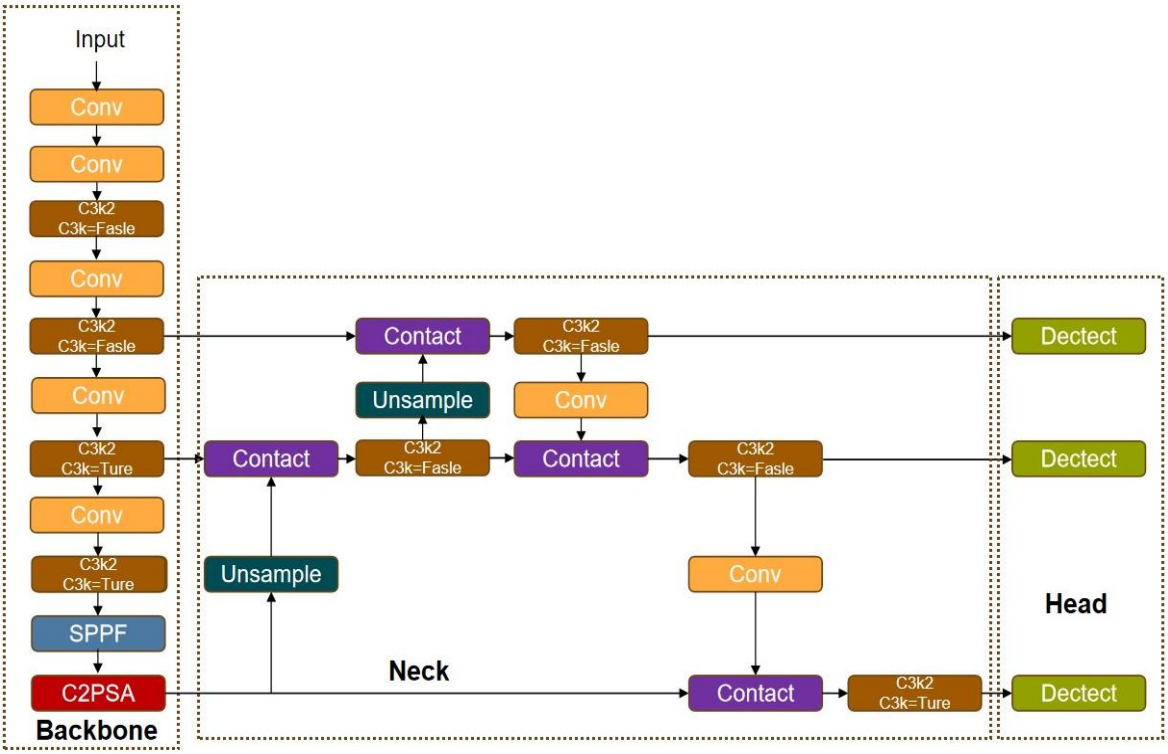

Figure 1 YOLOv11 structure diagram.

3. Improved models

As to address the limitations of the YOLOv11 model, a refined network model, YOLOv11-AAG, is introduced. The structure of the YOLOv11-AAG model is illustrated in Figure 2. Initially, the C3k2 module in the final layer of the backbone network is substituted with the C3k2-AKConv structure. Subsequently, the adaptive Multi-scale Feature Fusion Network (AMFFN)[14] is integrated into the neck section. The primary objective of AMFFN is to tackle challenges related to inadequate multi-scale feature fusion and the potential oversight of minor defects in metal defect identification. By leveraging spatial attention mechanisms and channel attention mechanisms, AMFFN can concentrate on small defect regions on steel surfaces, mitigate background noise, and amplify the representation of defect characteristics. Additionally, it can dynamically adjust the fusion weights of various feature levels, circumventing the constraints of conventional fixed fusion techniques and achieving more precise multi-scale feature fusion. Lastly, given the diverse geometries of metal defects, an Adaptive Defect Sensing Anchor Frame Generation Mechanism (Guided-Anchoring) is proposed. This mechanism enables precise adaptation to intricate metal defect shapes by dynamically modifying the size, aspect ratio, and distribution density of anchor

frames.

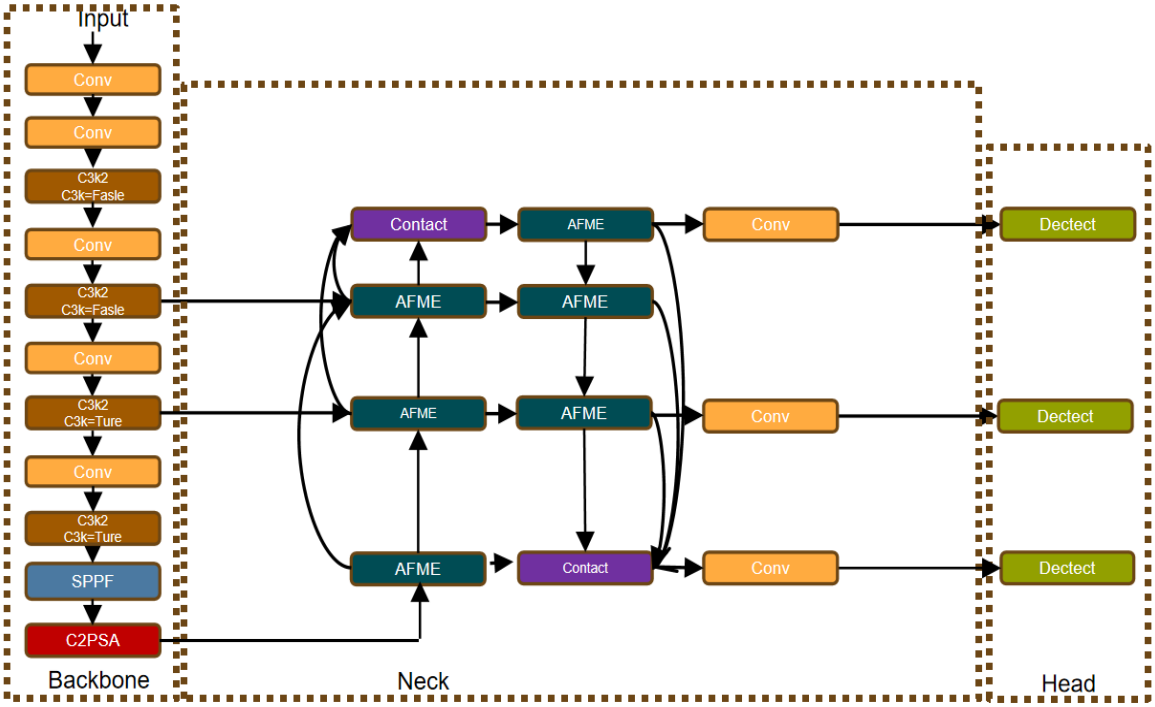

Figure 2 Structure diagram of YOLOv11-AAG.

3.1 C3k2-AKConv

YOLOv11 standard convolution square kernel structure has stable feature extraction ability for regular objects (such as rectangular welding area), which meets the basic modeling requirements of YOLO series for general objects, and its reasoning speed is fast and insensitive to noise. However, it has some shortcomings in metal defect detection, such as low efficiency of feature pyramid for corrosion area with fuzzy edg[15]; and YOLOv11 standard convolution square kernel can not adapt to complex geometric shapes such as zigzag crack, annular hole, irregular corrosion boundary, etc. The deformable convolution module AKConv can dynamically select the kernel size according to the input characteristics, and its single module can handle defects of different scales without stacking multilayer convolution. On the Severstal steel defect dataset, AKConv replaced the standard convolution in Backbone, resulting in an 88% improvement in Recall from 72% for small targets (cracks) and an 88% improvement in Precision from 65% to 76% for large targets (corrosion). AKConv is also able to accurately model irregular shapes for the diversity of metal defects by learning offsets to generate sampling meshes that match the shape of defects. For deformable convolution, assume that the input feature map $X \in R^{B\times Cin\times H\times W}$ represents the number of channels C, the height H, and the width W. It is a dynamic adaptation feature. The spatial variance of the feature map will be calculated first. The process is described by formula as shown in formula (1):

$$Var\ (X) = \frac{1}{B\times C_{in}\times H\times W}\sum_{b=1}^{B}\sum_{c=1}^{C_{in}}\sum_{i=1}^{H}\sum_{j=1}^{W}(X_{b,c,i,j}-\bar{X}) \quad (1)$$

The offset Δp is forecasted through convolution and integrated with conventional grid coordinates to derive offset sampling coordinates P. These coordinates are then subjected to bilinear interpolation to sample from the initial feature map. Subsequently, the resultant kernel parameters are employed to dynamically convolve the sampled feature map, as depicted in formula (2).

$$Y(b,c',i,j)=\sum_{c=1}^{C_{in}}\sum_{m=1}^{k}\sum_{n=1}^{k}\theta'_{k}(b,c',c,m,n)\cdot X_{sample}(b,c,i+m,j+n) \quad (2)$$

The output results are normalized and activated to achieve efficient detection of multi-scale and irregular targets. Its overall structure and flow are shown in Figure 3.

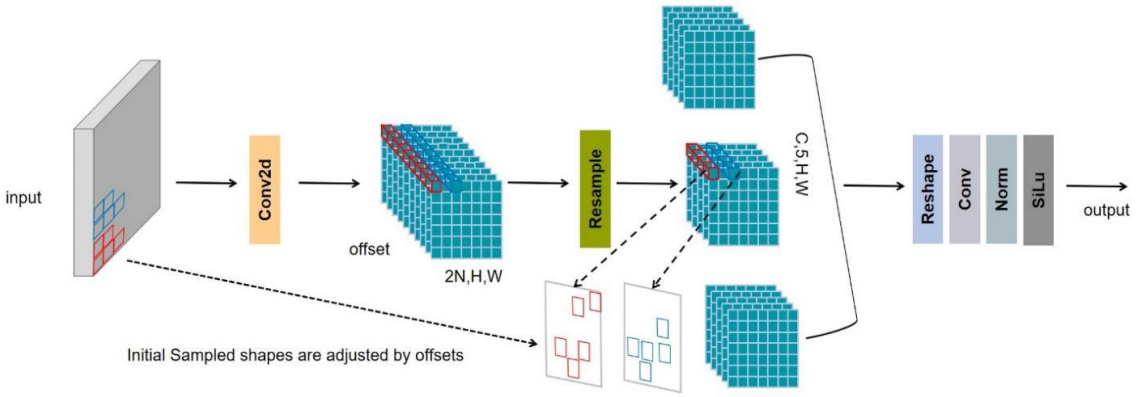

Figure 3 AKConv Structure and Flow Chart.

To effectively integrate the C3K2 module with AKConv[15], determining the optimal positioning of the combination is crucial. In this study, the C3K2 module in the sixth layer of the backbone network is refined into C3K2-AKConv. This modification markedly enhances the detection accuracy of irregular targets like metal defects while preserving real-time performance. The refined architecture is illustrated in Figure 4.

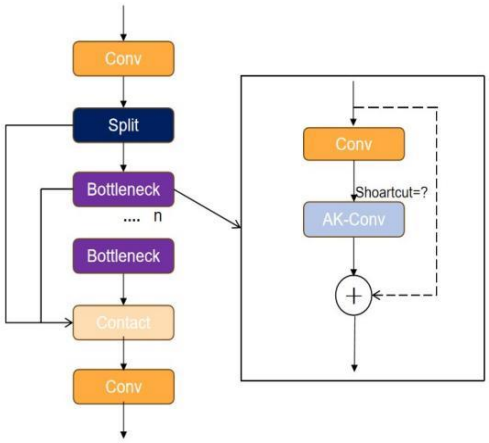

Figure 4 Structure of C3K2-AKConv.

3.2 AMFFN

In contrast to Bifpn's limitations in detecting metal defects, Amffn employs a trainable weight generator to dynamically compute fusion weights for various feature levels based on input feature characteristics such as scale and shape. This approach enhances the detection of micro cracks by augmenting the weight of low-level features while preserving high-resolution details. For large-area corrosion detection, Amffn boosts the semantic information of high-level features and utilizes a dynamic weight mechanism to address the constraints of fixed fusion[16]. Additionally, by integrating the edge enhancement convolution module AKConv, Amffn can concentrate on defect edges and textures to amplify defect specifics.

The adaptive multi-scale feature fusion network (AMFFN) comprises four primary components, delineated in Figure 5: original feature extraction, adaptive multi-scale feature extraction (AMFE), feature fusion, and image reconstruction[17].

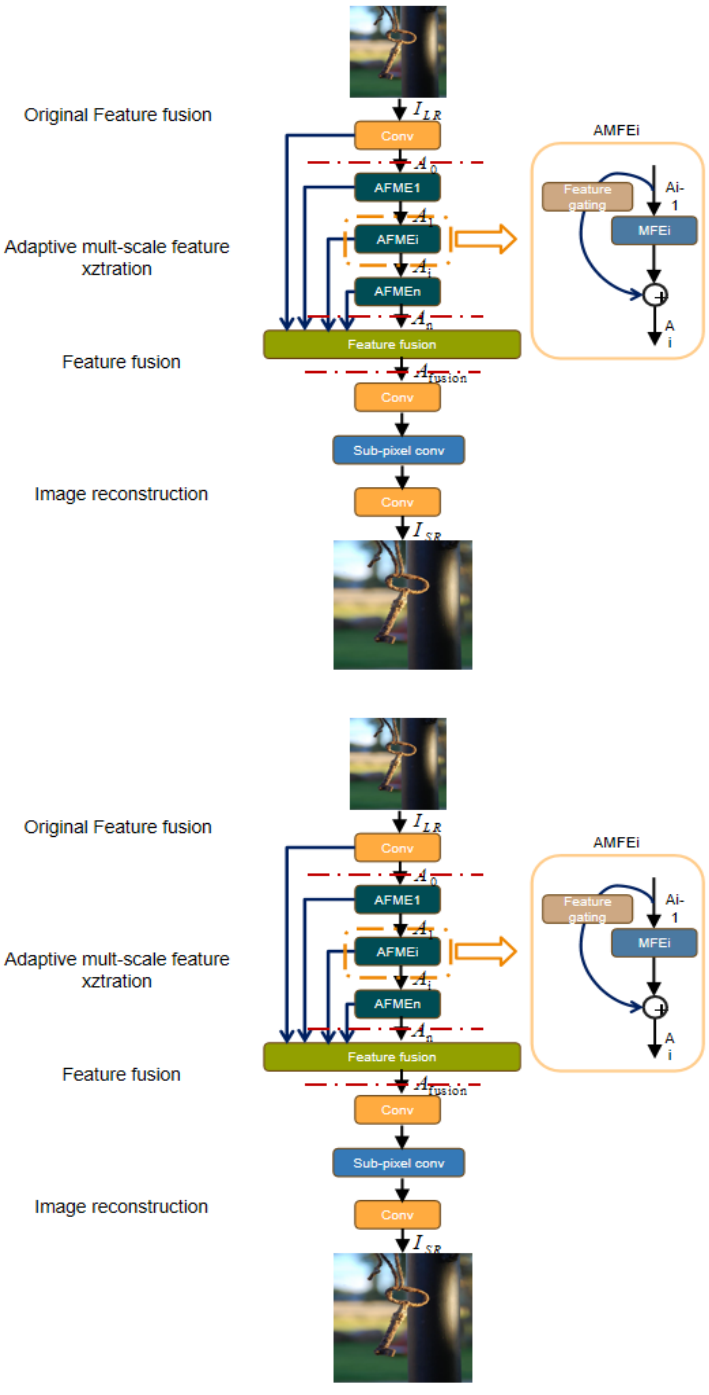

Figure 5 Network structure of adaptive multiscale feature fusion network (AMFFN).

First, convolution layer conv with n0 convolution kernels is applied to the input image to generate a set of feature maps, specifically as follows:

$$A_0 = \mathrm{w}_0 * I_{LR} + \mathrm{b}_0 \qquad (3)$$

Where A_0 is used as the original feature extracted from the low-resolution metal defect image, w_0 corresponds to the filter in the convolution layer, here 128 convolution kernels with spatial size of 3×3, b_0 represents the offset of the convolution layer, and ' *' represents the convolution operation. In the adaptive multi-scale feature extraction part, assuming that this part contains n adaptive multi-scale feature extraction blocks (AMFE), then the ith AMFE can be expressed as:

$$A_i = \mathrm{f}_{MFE}(A_{i-1}) + g(A_{i-1}) \quad (1 \le i \le n) \tag{4}$$

These feature maps contain a lot of redundant information, which will greatly increase the computational burden if they are directly used for image reconstruction. Therefore, before inputting these features for super-resolution into the reconstruction layer, the feature fusion layer is set after n AMFE for feature fusion and dimensionality reduction. The output formula of the feature fusion layer A_{fusion} is:

$$A_{fusion} = w_f * [A_0, A_1, \cdots, A_0] + b_f \tag{5}$$

w_f corresponds to the weight of the feature fusion layer, representing 64 convolution kernels with a size of 1× 1, b_f is the corresponding deviation, $[A_0, A_1, \cdots, A_n]$ represents the parallel connection of all feature maps extracted by the first feature extraction layer and AMFE.

Adaptive Multiscale Feature Extraction (AMFE) module is the core module of AMFFN[18]. The structure of AMFE is shown in Figure 6.

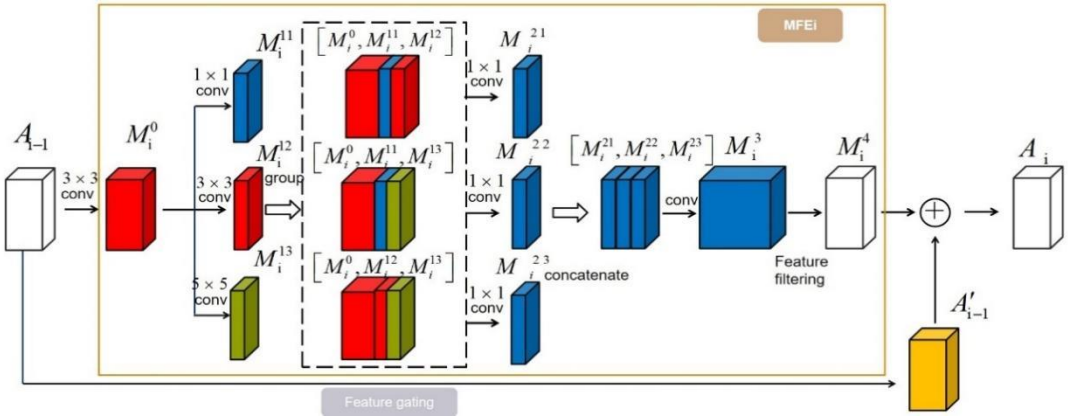

Figure 6 AMFE Module Details.

AMFE module firstly receives multi-scale feature map from backbone network, then generates spatial attention map M_S through AKConv convolution and global pooling, focuses defect area and suppresses background, and calculates channel attention weight by global average pooling and full connection layer to screen valuable feature channels for defect detection; After multiplying the spatial and channel attention maps with the input features to obtain enhanced features F', dynamic weights are generated by calculating feature statistics and small neural networks.

3.3 Adaptive defect sensing detector

Due to the constraints of the YOLOV11 detection head's fixed frame and its limited adaptability to defects, this study employs an alternative approach called Guided-Anchoring, as illustrated in Figure 7, to address the challenges in metal defect detection.

An anchor generation module with dual branches is employed for each output feature map within

the pyramid to predict anchor positions and shapes. Subsequently, a feature adaptation module is applied to the original feature map to enable the new feature map to recognize anchor shapes.

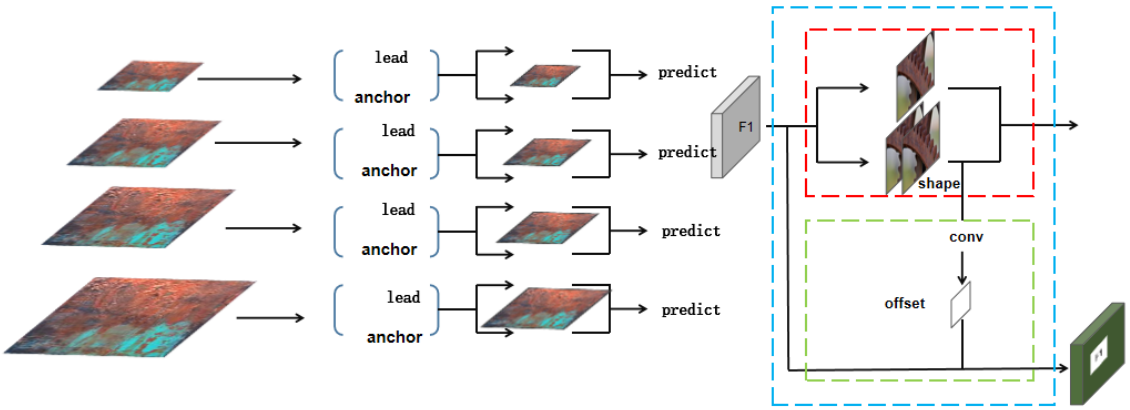

Figure 7 Guided-anchoring framework.

In the actual working process, the improved detection head first uses the defect feature module to further process the feature map, and at the same time, outputs the probability distribution ρ that different regions in the feature map belong to various defects through the defect category prediction branch, and then obtains the spatial attention weight α through the spatial attention calculation branch; then uses the dynamic anchor frame generator module to generate and adjust the width and height of the anchor frame, and uses the formula to describe as shown in Formulas (9) and (10):

$$\alpha_{\omega} = \alpha_{\omega_{\text{init}}} \times \left(1 + \lambda_{\omega} \sum \rho_{(c_i)} \cdot \omega_i\right) \tag{6}$$

$$\alpha_{h} = \alpha_{h_{\text{init}}} \times \left(1 + \lambda_{h} \sum \rho_{(c_i)} \cdot h_i\right) \tag{7}$$

Where $\alpha_{\omega_{\text{init}}}$ and $\alpha_{h_{\text{init}}}$ are the width and height of the initial anchor box, λ_{ω} and λ_{h} are scaling coefficients, w_i and h_i are the width and height of the corresponding cluster center.

4. Experiments and results

4.1 Experimental environment

The experimental setup utilized a computer system running Windows 11 64-bit, equipped with an Intel i7-11800H processor, an NVIDIA GeForce RTX 306 graphics card, operating at a base frequency of 2.3GHz, and 16GB of RAM. The system also featured CUDA version 11.8, utilized Python 3.10 as the programming language, and employed PyTorch 2.0.1 as the deep learning framework.

The initial learning rate, batch size, number of rounds, input image size, weight attenuation coefficient, and IoU threshold are specified as follows: .01, 8, 200, 640×640, .00005, and .5, respectively.

4.2 Data sets and preprocessing

The scarcity of existing research in this domain necessitated the creation of proprietary datasets for this study. A total of 2838 images depicting diverse steel defects were amassed from various industrial settings, categorized as corrosion, welding imperfections, holes, and cracks. We illustrate selected examples of these metal anomalies in Figure 8.

Figure 8 Examples of metal defects.

The dataset predominantly comprises metal defect images from industrial settings, with images resized accordingly. A total of 5676 images were augmented and annotated. The augmented dataset was partitioned into training, testing, and validation sets in a 7:2:1 ratio.Training was conducted using the PyTorch framework with the YOLOv11 base model and YOLOv11-AAG. The training process involved 200 epochs, with training halted if no improvement in parameters was observed after 20 consecutive epochs.

4.3 Evaluation index

To enhance the evaluation of the refined model, the following metrics were chosen: Recall (R), Precision (P), Mean Average Precision (mAP), Parameters (Params), Gigaflops (GFLOPs), Anchor Frame Coverage of Defects, Average Recall at 100 detections (AR@100), and Frames Per Second (FPS).

4.4 Ablation experiments

Table 1 presents the results of the ablation experiments on the improvements made to the original YOLOv11 model in this study.

Table 1 Experimental results of ablation with different modifications of YOLOv11.

AKConv	AMFFN	Guided-	P/%	R/%	mAP/%	Weight/MB
-	-	-	81.4-	71.1	77.1	24.2
√			81.2	72.4	77.8	24.7
	√		82.7	75.7	78.6	26.8
		√	81.9	72.0	77.7	24.6
√	√		83.7	75.3	80.1	27.1
	√	√	83.8	75.2	79.4	27.3
√		√	82.5	72.8	78.7	25.9
√	√	√	82.3	76.8	80.3	27.7

4.5 Comparative Experimental Analysis of Different Models

To validate the efficacy of the YOLOv11-AAG model proposed in this study, we compared it with several well-known convolutional neural network models, including YOLOv5, YOLOv8, SSD, Faster R-CNN, and YOLOv11. Training and testing were conducted on a proprietary dataset of metal defects, yielding the performance outcomes for each network as detailed in Table 2

Table 2 Comparison results of defect detection among different models

model	P/%	R/%	mAP/%	Weighted documents/MB
YOLOv5	67.6	56.5	68.2	5.4
YOLOv8	77.0	65.4	73.8	8.1
SSD	67.5	60.5	67.9	67.4
Faster R-CNN	54.3	63.2	56.5	371.2
YOLOv11	81.4	71.1	77.1	21.2
YOLOv11- AAG	82.3	76.8	80.3	27.7

4.6 Verification of experimental detection effect

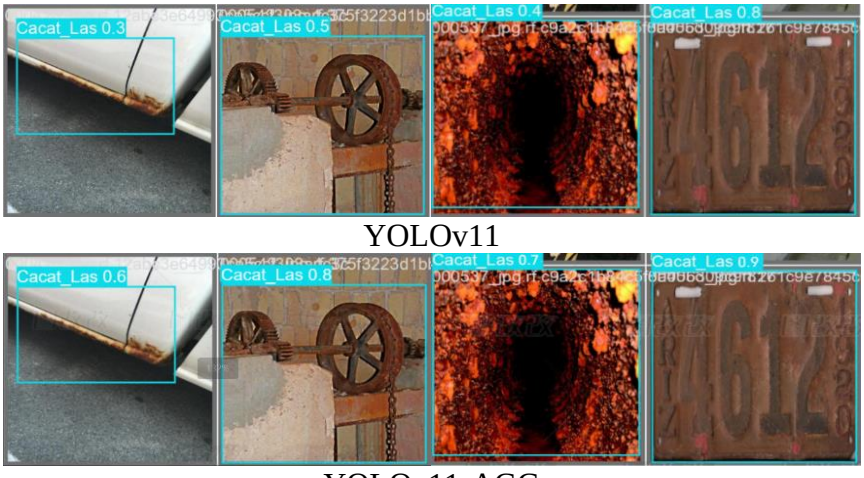

Figure 9 Comparison of detection effect.

Figure 9 show the detection results of the YOLOv11-AGG model. It can be observed from these four groups of images that, compared with the original model, the improved model proposed in this paper can maintain higher confidence when detecting seam heads. Additionally, the range of detection boxes is reduced in some images. All experimental results demonstrate that the improved algorithm model based on yolov11 proposed in this paper can achieve more effective detection of fabric seam heads in complex industrial environments, while maintaining extremely high accuracy and thus exhibiting superior application value.

5. Conclusions

A refined YOLOv11-AAG model is proposed for detecting metal defects, building upon the original YOLOv11 architecture. Experimental results demonstrate that YOLOv11-AAG outperforms YOLOv8 and Faster R-CNN in complex industrial settings, achieving a detection accuracy increase from 77.1% to 80.3%. The model exhibits enhanced capability in identifying fine defects like cracks and holes while reducing computational complexity through structural

optimization. This optimization balances detection efficiency and accuracy, catering to real-time detection requirements on metal production lines. The proposed model offers an effective solution for automating quality assessment in metal product manufacturing, thereby advancing industrial intelligent detection technology.

References

[1] Zhang Wei, Yuan Gaihong, Zhou Tianyi, et al. Review of metal surface defect detection methods for aerospace equipment based on machine learning [J]. Aerospace Electronic Countermeasures, 2025, 41 (03):43-50.

[2] Li Tengyue. Research on surface defect detection algorithm of trigeminal axis based on TS-YOLO [D]. Zhejiang University of Science and Technology, 2024.

[3] Xu Yedong, Cai Yaheng, Li Yan et al. Lightweight overhead transmission line nest detection network based on YOLOv5s [J]. Electronic Measurement Technology, 2024, 47 (7):138-148.

[4] DONG CH, ZHANG K, XIE ZH Y, et al. An improved cascade RCNN detection method for key components and defects of transmission lines[J]. IET Generation, Transmission Distribution, 2023, 17(19): 4277-4292.

[5] Liang D, Ye JM, Zhao K, et al. Machine vision inspection of bearing defects based on data enhancement and SSD [J]. Mechanical Manufacturing, 2025, 63 (04):18-23.

[6] Xu Changdong. Research on Fan Blade Defect Detection Based on Improved YOLO Algorithm [D]. Hubei University for Nationalities, 2025.

[7] Chen Xi. Object Recognition Based on Micro-network Architecture Search [D]. North China University of Technology, 2022.

[8] Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection[J]. Machines, 2023, 11(7): 677.

[9] Yang Hang. Research on speed limit sign recognition technology based on machine visi on [D]. Wuhan University of Technology, 2019.

[10] Zhang Xin. Research on UAV Image Target Detection Based on Improved Convolution al Neural Network [D]. Chongqing Normal University, 2024.

[11] Ming Li, Hao Su, Jia Cheng. CFAI-YOLO: an enhanced pedestrian and vehicle detection algorithm based on YOLOv11s[J/OL].Optoelectronics Letters,1-9[2025-08-10].

[12] Tang Feng. Research and Optimization of Industrial Defect Detection Algorithm Based on Deep Learning [D]. University of Electronic Science and Technology, 2025.

[13] Li Hang. Research on multi-scale target detection method based on anchorless frame [D]. Guilin University of Electronic Technology, 2024.

[14] Wang Xinying. Remote sensing image super-resolution reconstruction method based on multi-scale feature adaptive fusion network [D]. Hubei University of Technology, 2020.

[15] Lei Xin. Digital signal processing of lung medical images based on deep learning [D]. University of Electronic Science and Technology, 2024.

[16] Shen H J, Li H Y, Huang Y P, et al. Vehicle detection method based on adaptive mu lti-scale feature fusion network [J/OL]. Journal of Electronics, 1 -9[2025-08-10].

[17] Thammasanya T, Patiam S, Rodcharoen E, et al. A new approach to classifying polym er type of microplastics based on Faster-RCNN-FPN and spectroscopic imagery under ultraviolet light[J]. Scientific reports, 2024, 14(1): 3529.

[18] Tian Wei Zhou, Yu Luo, Guan Hui Yue, et al. Adaptive Mixed-Scale Feature Fusion Network for Blind AI-Generated Image Quality Assessment [J]. IEEE Transactions on Broadcasting, 2024, 70 (3): 1328-1339.