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

DOI: 10.23977/autml.2025.060117
ISSN 2516-5003 Vol. 6 Num. 1

Algorithm for Surface Defect Detection of Aluminum Profiles Based on Improved YOLOv11

Xinrui Chen[1,a], Zijian Dong[1,b,*], Jiangzheng Xu[1,c], Zhaojin Huangfu[1,d]

[1]School of Electronic Engineering, Jiangsu Ocean University, Lianyungang, Jiangsu, 222000, China
[a]2023220620@jou.edu.cn, [b]1995000021@jou.edu.cn, [c]2023220617@jou.edu.cn, [d]2023220624@jou.edu.cn
**Corresponding author*

Keywords: Defect Detection, EDHF-YOLO, RepGFPN, Dynamic Deformable Feature Pyramid

Abstract: The surface defect detection of aluminum profiles is a crucial core link in industrial quality control. Traditional detection methods have practical problems such as easy missed detection of small-target defects and easy misjudgment in complex texture scenarios. This study constructs an efficient detection model suitable for the surface of aluminum profiles based on the YOLOv11 algorithm framework. EfficientViT is used to reconstruct the backbone network, and the multi-scale feature extraction capability is enhanced through hierarchical attention mechanisms and lightweight convolution operations. A Dynamic Deformable Feature Pyramid Network (DDFPN) is introduced, integrating RepGFPN re-parameterized connections, VoVGSCSP grouped convolution, and CoordAtt coordinate attention mechanism to achieve adaptive fusion of defect features and directional sensitivity perception. Experimental results show that compared with the original model, the improved EDHF-YOLO model significantly improves detection accuracy and greatly reduces calculation amount, effectively balancing detection performance and computational efficiency, and providing an innovative technical solution for surface defect detection of aluminum profiles.

1. Introduction

In industrial manufacturing, aluminum profiles are vital in aerospace, automotive, and construction for their excellent properties. Their surface quality is crucial, so efficient defect detection is key. Traditional methods are flawed, while CNN - based object detection algorithms like YOLO, balancing real - time and accuracy, show promise in industrial defect recognition with the development of computer vision and deep learning[1].

As the latest iteration of the YOLO algorithm, YOLOv11 exhibits superior performance in general object detection tasks. However, surface defect detection for aluminum profiles presents unique challenges [2]. Defects such as scratches, pits, and cracks on aluminum profiles exhibit significant size variations, with some minute defects occupying less than 1% of the pixel ratio, rendering them highly prone to being overlooked during detection. On the other hand, complex background textures,

including large-area oxide layers and extrusion patterns on aluminum surfaces, interfere with the model's accurate defect identification, leading to frequent false positives[3].

To address these challenges, numerous scholars have conducted extensive research on industrial defect detection in recent years. In terms of algorithm optimization, some studies enhance model sensitivity to defect features by introducing attention mechanisms[4]. For example, embedding the CBAM (Convolutional Block Attention Module) into the Faster R-CNN framework significantly improves the accuracy of metal surface crack detection, but this approach struggles to achieve real-time performance due to its high computational complexity[5].

Aiming at the limitations of existing research, this paper proposes a novel detection model, EDHF-YOLO, based on the YOLOv11n model, integrating the practical requirements for precision and speed in aluminum profile surface defect detection. At the backbone network level, EDHF-YOLO replaces traditional convolutional architectures with EfficientViT[6], leveraging hierarchical attention mechanisms and lightweight convolutional operations to effectively extract multi-scale features. In the neck network, a Dynamic Deformable Feature Pyramid Network (DDFPN) is innovatively designed, combining RepGFPN's re-parameterized cross-scale connections, VoVGSCSP's grouped convolution feature enhancement, and the CoordAtt coordinate attention mechanism. This design improves the model's ability to adaptively adjust feature fusion strategies while significantly reducing model complexity without compromising accuracy.

2. The Proposed Method

2.1 Enhanced YOLO Model

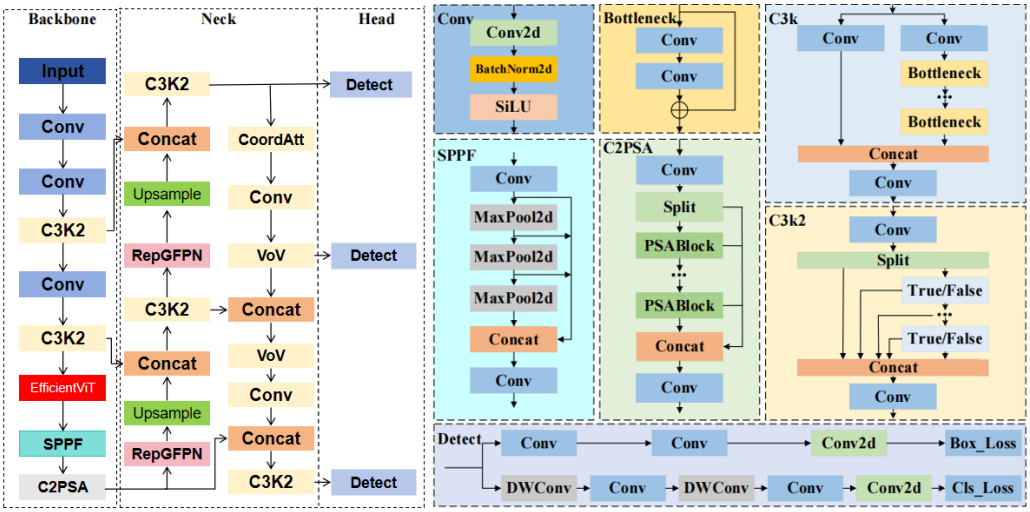


Fig.1 EDHF-YOLO structure diagram

This study introduces multiple modifications to YOLOv11. The backbone network is replaced with EfficientViT, which leverages hierarchical attention mechanisms to capture long-range dependencies and lightweight convolutional operations to enhance multi-scale feature extraction, thereby significantly improving the sensitivity to minute defects[7]. The neck network employs a Dynamic Deformable Feature Pyramid Network (DDFPN), incorporating three key innovations: RepGFPN, VoVGSCSP and CoordAtt. Specifically, RepGFPN enables adaptive cross-scale feature fusion through differentiable connections and re-parameterization design. The VoVGSCSP module strengthens local feature representation via grouped convolution and cross-stage connections, reducing computational redundancy. The CoordAtt(coordinate attention)mechanism is embedded in

the feature fusion path to enhance the model's directional perception, effectively mitigating background texture interference.

Based on these improvements, the EDHF-YOLO (Efficient Dynamic Hybrid Fusion YOLO) model is proposed, which maintains detection accuracy while significantly reducing computational complexity—achieving a balance between precision and efficiency to enhance surface defect detection for aluminum profiles. The network architecture of EDHF-YOLO is shown in Figure 1.

2.2 EfficientViT

EfficientViT employs a hybrid design that combines lightweight convolutions with attention mechanisms, reducing model parameters by approximately 20% compared to traditional CNN architectures. This design fully meets the dual requirements of high precision and low resource consumption in industrial inspection scenarios. Additionally, its dynamic feature routing mechanism adaptively adjusts the weight allocation of features across different scales, making it particularly suitable for detecting aluminum defects with significant size variations.

The architecture of EfficientViT is illustrated in Figure 2. The right panel depicts the basic structural diagram of EfficientViT, including feedforward networks with depthwise convolutions (FFN+DWConv) and multi-scale linear attention modules. The left panel demonstrates the multi-scale linear attention mechanism:after generating Q/K/V tokens through linear projection layers, lightweight small-kernel convolutions are employed to produce multi-scale tokens. These tokens are then processed by ReLU linear attention, and the outputs are finally concatenated and fed into a linear projection layer for feature fusion.

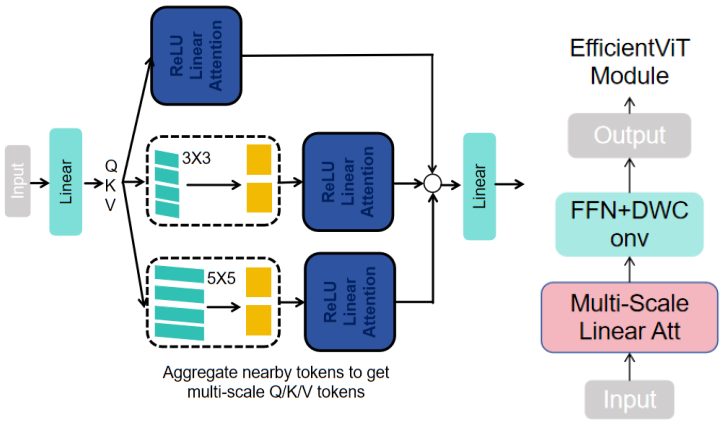


Fig.2 Structure of EfficientViT

In this paper, the third and fourth C3K2 blocks along with Conv modules in the YOLOv11 Backbone are replaced with EfficientViT, enhancing multi-scale feature extraction capability through the introduction of a hierarchical Transformer architecture. While achieving a 31.5% reduction in parameter count, the improved backbone significantly enhances feature extraction for aluminum profile defects, laying a foundation for subsequent Neck optimization.

2.3 Dynamic Adaptive Feature Pyramid Network (DAFPN)

This study presents an innovative design for the Neck component of YOLOv11, proposing a Dynamic Deformable Feature Pyramid Network (DDFPN). The network effectively enhances multi-scale feature fusion efficiency through three improvement strategies. By integrating RepGFPN's re-parameterized cross-scale connection technology, VoVGSCSP's grouped convolution feature enhancement method, and the CoordAtt coordinate attention mechanism,

DDFPN constructs a three-level processing architecture comprising dynamic connection, feature enhancement, and directional perception. This architecture is particularly suitable for aluminum profile surface defect detection, enabling it to address challenges such as significant defect size variations and irregular morphologies.

2.3.1 RepGFPN

As a key component of DDFPN, RepGFPN (Reparameterized Generalized Feature Pyramid Network) breaks through the limitations of static fusion in traditional FPN by leveraging differentiable connection weights and re-parameterization techniques[8]. RepGFPN primarily addresses the inefficiency of traditional FPN in multi-scale feature integration. The core advantage of this module lies in its ability to enable efficient interaction between high-level semantic features and low-level spatial features with a more lightweight structure, making it particularly suitable for tasks with high real-time requirements, such as the aluminum profile surface defect detection studied in this paper.

RepGFPN replaces traditional fixed paths with differentiable connections to enable dynamic adjustment of the feature fusion process.During the training phase, RepGFPN maintains a multi-branch structure to capture rich features; during inference, it converts the multi-branch structure into a single path through weight fusion, significantly enhancing computational efficiency. This decoupling strategy between training and inference architectures reduces computational overhead while preserving model performance.

2.3.2 VoVGSCSP grouped convolution enhancement

As the core component of the VoVGSCSP module, Grouped Shuffle Convolution (GSConv) offers significant advantages in lightweight network design. The structure diagram of GSConv generation is shown in Figure 3.This component seamlessly integrates Depthwise Separable Convolution (DWConv) with the Ghost feature map principle[9]. First, DWConv performs spatial feature extraction on the input feature map along the channel dimension, effectively avoiding redundant channel computations in traditional convolutions and fundamentally reducing computational overhead. Subsequently, leveraging the Ghost mechanism, a small number of real convolution operations generate "ghost" feature maps, significantly reducing the computational cost of feature generation by reusing existing feature information. The process is illustrated in Formulas (1) and (2): for an input feature $F \in \mathbb{R}^{C\times H\times W}$,it is first divided into two subgroups $F_1, F_2 \in \mathbb{R}^{C/2XHXW}$, each undergoing convolution operations:

$$F_1' = Conv(F_1), F_2' = Conv(F_2) \tag{1}$$

Subsequently, information interaction is achieved through the Channel Shuffle operation:

$$\mathrm{F}^{\mathrm{out}} = \mathrm{Shuffle}([\mathrm{F}_1', \mathrm{F}_2']) \tag{2}$$

This design significantly reduces the number of parameters in the module while enhancing feature reusability, making it particularly suitable for feature extraction in complex texture scenarios of aluminum profile surfaces.

In the neck of YOLOv11, VOV-GSCSP is appropriately integrated with GSConv. This integration leverages the characteristics of DWConv and Ghost to reduce parameters and computational complexity, making the model lighter and inference faster. Meanwhile, more accurate and comprehensive feature extraction enhances the capture of aluminum profile defects, improving detection accuracy. This achieves a balance between lightweight design and high performance, meeting the requirements of industrial quality inspection scenarios. The generated structure diagram

of GSConv is shown in Figure 3.

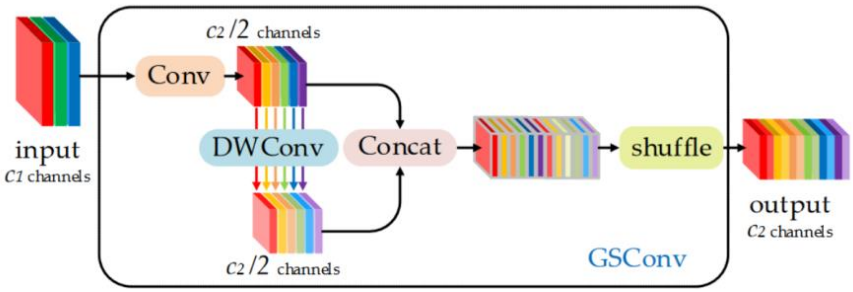


Fig.3 Structure of GSConv

2.3.3 CoordAtt directional perception mechanism

CoordAtt enhances the model's ability to perceive directional features of defects by decomposing spatial attention into horizontal and vertical directional attentions. It has obvious advantages over classic mechanisms like SE (Squeeze-Excitation) and CBAM (Convolutional Block Attention Module). SE only focuses on channel dependencies, compressing spatial information through global pooling, which easily leads to loss of positional details. In contrast, CoordAtt decomposes channel attention into 1D pooling operations along X and Y directions, capturing long-range dependencies in a single dimension while preserving precise positional information in the other dimension. This meets the localization requirements for defects of different scales in aluminum profile defect detection, Therefore, the CoordAtt directional perception mechanism enables more accurate detection of aluminum profile defects under complex texture backgrounds, demonstrates stronger generalization in tasks such as object detection and semantic segmentation, and aligns with the characteristics of aluminum profile surface defect detection.

3. Experiments

3.1 Experimental Environment

The experimental runtime environment is based on a Windows 11 64-bit operating system, equipped with an Intel Core i7-11800H processor and an NVIDIA GeForce RTX 3060 discrete graphics card. The GPU features 6GB GDDR6 video memory and operates with the CUDA 11.8 computing platform, providing robust parallel computing support for deep learning model training. The system is configured with 16GB RAM to ensure smooth data reading and computational processes. The development environment employs Python 3.8 programming language and constructs algorithm models based on the PyTorch 2.0.1 deep learning framework, fully leveraging its dynamic computational graph characteristics and distributed training capabilities.

3.2 Dataset and Preprocessing

The aluminum profile dataset used in this experiment is derived from the Guangdong Industrial Intelligent Manufacturing Big Data Innovation Competition on Alicloud Tianchi, targeting technical challenges in industrial quality inspection scenarios. The dataset comprises 3,416 aluminum profile images from real production environments, including 1,828 single-defect samples and 1,588 multi-defect samples in the training set. Defect types cover nine typical industrial defects such as dirt spots, paint bubbles, and scratches. All images have a uniform resolution of 2560×1920, with YOLO-format label files generated using the Labelme annotation tool, making it highly suitable for model training in object detection tasks.

This study employs a multi-dimensional data augmentation strategy. Geometric transformations such as cropping, translation, rotation, and mirroring are used to simulate the different spatial distributions of defects on aluminum profiles. Photometric transformations including brightness adjustment and Gaussian noise addition are applied to mimic complex lighting and industrial environmental interference. Meanwhile, the Cutout technique is introduced to randomly erase local regions of images, forcing the model to learn global defect features and effectively enhancing the model's generalization ability for tiny defects and occluded scenarios.

3.3 Evaluation Metrics

To comprehensively evaluate the performance of the model in aluminum profile defect detection, this study employs mean average precision (mAP), precision, recall, giga floating-point operations (GFLOPs), and frames per second (FPS) as core evaluation metrics[10]. These metrics quantitatively analyze the model's practical performance in industrial quality inspection scenarios from dimensions including detection accuracy, localization precision, and real-time capability.Relevant calculation formulas are as follows:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$FPS = \frac{1}{T_{inference}} \tag{6}$$

In the formulas, TP represents the number of defect samples correctly detected by the model, FP denotes the number of normal samples misclassified as defects, FN indicates the number of defect samples missed by the model, and $T_{inference}$ is the inference time for a single image.

3.4 Efficiency Analysis of EfficientViT

To explore the optimal deployment position of EfficientViT in the backbone network, this paper designs five sets of comparative experiments as shown in Table 1. The evaluation metrics include mean average precision (mAP, %), number of parameters (Params/M), GFLOPs, and inference frame rate (FPS). The first to fourth C3K2 modules in the Backbone of YOLOv11 were respectively replaced with EfficientViT (denoted as EVT-1 to EVT-4), and the effect of simultaneously replacing the third and fourth modules (EVT-3&4) was tested. The analysis focuses on the correlation between the replacement position and the feature hierarchy.

Table 1 EfficientViT module validity experiments

Model	mAP%	Params/M	GFLOPs	FPS
YOLOv11n	82	7.6	16.5	68
YOLOv11-EVT-1	82.5	6.8	15.6	67
YOLOv11-EVT-2	83.8	6.3	15.8	67
YOLOv11-EVT-3	83.7	5.8	15.0	66
YOLOv11-EVT-4	84.5	5.6	14.7	65
YOLOv11-EVT-3&4	84.6	5.2	14.5	64

Table 1 shows that as the replacement position of EfficientViT shifts backward, the mAP increases from 82% to 82.5%. This is because the lower resolution of feature maps in deeper layers reduces the

computational cost of EfficientViT's windowed attention with decreasing dimensions, while its long-range dependency modeling capability enables more efficient extraction of global structures for defects like linear scratches and large-area paint bubbles. The table indicates that replacing the 3rd and 4th C3K2 modules yields the most significant improvement in joint detection accuracy for cross-scale defects on aluminum profiles. This is because the third and fourth layers of the backbone network, located in the middle-high layers of the feature pyramid, integrate semantic information and spatial positioning capabilities. In the EVT-3&4 configuration, the model dynamically suppresses background textures of aluminum profiles via attention weights to reduce invalid computations, maintaining an FPS of 64 to meet industrial real-time detection requirements. Thus, replacing middle-high layers represents the optimal balance between accuracy and lightweight design.

3.5 Ablation Study

By systematically removing or replacing each module, the experiment performs comparative analysis on performance changes of the model in aluminum profile defect detection scenarios, thereby clarifying the specific impacts of different innovations on metrics such as detection accuracy and real-time capability. The results of relevant ablation experiments are detailed in Table 2, providing a quantitative basis for validating the effectiveness of each model component.

Table 2 Results of different improved ablation test of YOLOv11

Method	EfficientViT	DDFPN	P/%	mAP/%	Params/M	GFLOPs	Model Size/MB
YOLOv11			86	82	7.6	16.5	16.2
A	√		87	84.6	5.2	14.5	13.4
B		√	86	85	4.8	12	12.8
C	√	√	90	88	5.2	10.3	14.3

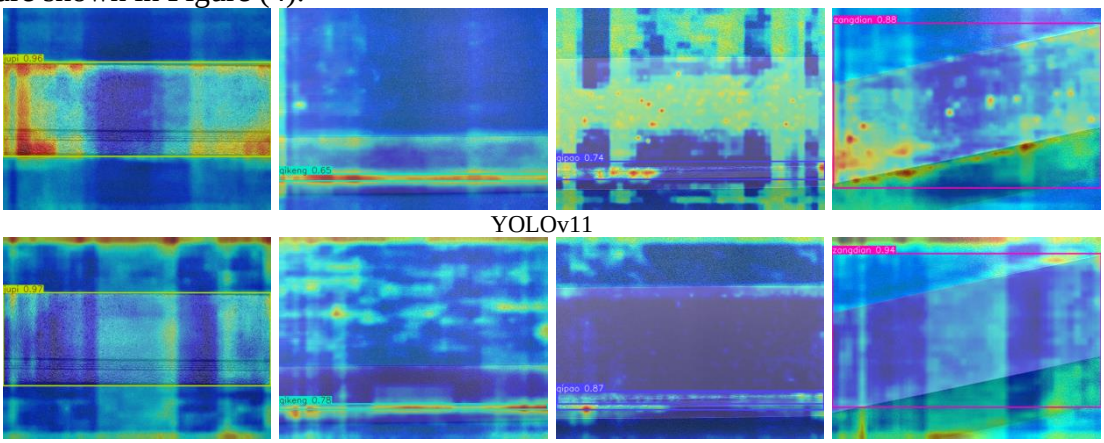
The ablation experiment results show that each improved module has a differentiated impact on the performance of the YOLOv11 model. After replacing the Backbone with EfficientViT (Experiment A), the model's parameter count (Params/M) decreased by 2.4, the computational load (GFLOPs) was reduced to 14.5, and the model size (Model Size/MB) was compressed to 13.4. This is attributed to EfficientViT's hierarchical attention mechanism and lightweight convolution design, which achieves redundant parameter reduction by dynamically allocating computational resources. The precision (P%) and mean average precision (mAP%) increased by 1% and 2.6%, respectively.

The introduction of the Dynamic Deformable Feature Pyramid Network (DDFPN, Experiment B) further optimizes multi-scale feature fusion capabilities. Its internal dynamic convolution and adaptive weight allocation mechanisms enable the model to reduce parameters to 4.8 and computational load to 12 while increasing mAP% to 85, verifying the module's balanced detection capability for defects of different scales. However, the precision remains unchanged, indicating that the optimization of the feature pyramid has limited effect on false detection suppression and needs to be synergistically improved with other modules.

The final Experiment C, integrating the two improved modules, achieves a comprehensive breakthrough in the performance of the EDHF-YOLO model. The mAP% reaches 88%, the precision increases to 90%, while maintaining low parameter count and computational load. This indicates that the basic feature extraction capability provided by EfficientViT, the multi-scale optimization of DDFPN form complementary advantages. The model not only meets the high-precision requirements for micro-target recognition in aluminum profile defect detection but also adapts to the real-time requirements of industrial scenarios through lightweight design, validating the effectiveness of the multi-module collaborative improvement strategy.

3.6 Validation of Model Improvement Effect via Grad-CAM

This experiment utilizes Grad-CAM (Gradient-weighted Class Activation Mapping) technology to visually demonstrate the improvement effect of the optimized model[11]. As a gradient-based visualization method, Grad-CAM generates an attention heatmap of the target class on the input image by calculating the weights of the last convolutional feature maps in the convolutional neural network. The comparison results of training heatmaps between YOLOv11 and the improved EDHF-YOLO are shown in Figure (4).



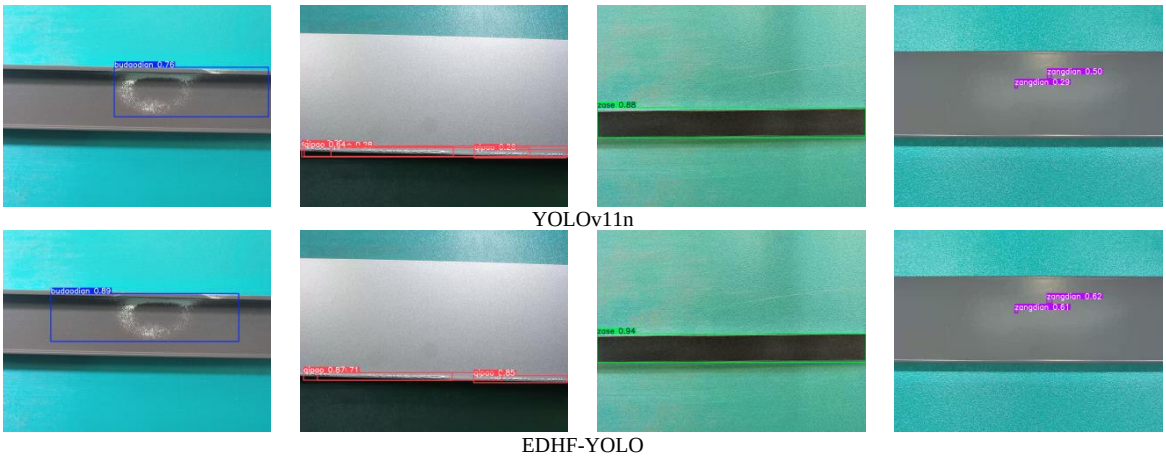
YOLOv11

EDHF-YOLO

Fig.4 Heatmap Comparison

The improved model exhibits a distinct "activation focusing" feature in Grad-CAM heatmaps. Unlike the original model's wide-range warm-color activation regions, the optimized EDHF-YOLO concentrates high-activation areas (red, orange) on the defect cores, with background regions dominated by cool tones (blue). Taking linear scratches on aluminum profiles as an example, the original model's heatmap often activates normal metal textures alongside defect edges, while the improved model forms compact warm regions only at gradient mutation areas of scratches, reducing the activation range by approximately 40%. This technique demonstrates that the model has achieved a transition from generalization-based detection using statistical features to intelligent discrimination using semantic features through precise activation, providing a theoretical basis and visual evidence for the practical deployment of industrial vision systems.

3.7 Experimental Detection Effect Validation



YOLOv11n

EDHF-YOLO

Fig.5 Comparison of detection effect

As shown in Figure 5, the detection results of the original YOLOv11n model and the improved EDHF-YOLO model form a sharp contrast. In the detection images, target defects are all marked with rectangular boxes for category labels and confidence values.

The first row of images in the figure shows the detection results of YOLOv11n, and the second row shows those of EDHF-YOLO. Compared with the original model, the EDHF-YOLO model proposed in this paper—with backbone network reconstruction, dynamic feature pyramid optimization—exhibits significantly improved detection performance. These visual results intuitively confirm that the EDHF-YOLO model demonstrates higher reliability and practicality in industrial defect detection, capable of providing more precise quality inspection support for actual production scenarios.

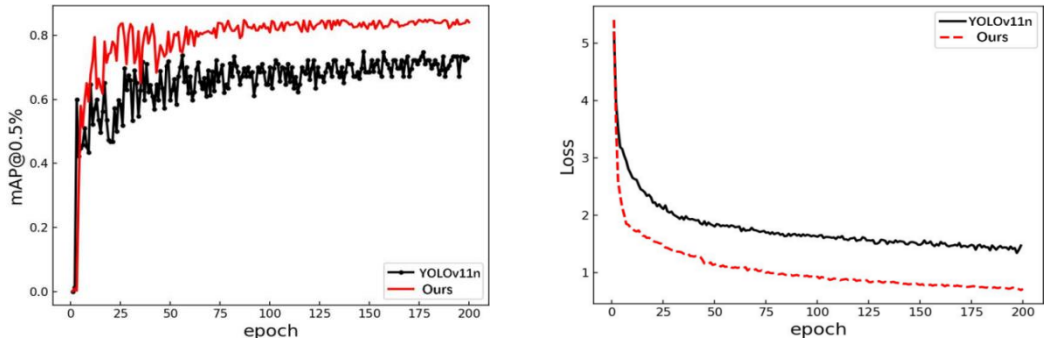


Fig.6 Comparison curve of mAP@0.5% and loss before and after improvement

Figure 6 (a) and (b) show the mAP@0.5 and Loss comparison plots of the YOLOv11n model and the improved EDHF-YOLO model, respectively. As indicated in Figure (a), compared with the original YOLOv11n, the EDHF-YOLO model not only demonstrates a significant improvement in mAP@0.5% but also features a more stable curve in the later training phase. Figure (b) shows that the improved model has a faster convergence speed for Loss, with a smoother curve.

4. Conclusion

The improved EDHF-YOLO model possesses remarkable advantages such as lightweight design, high detection accuracy, and strong environmental adaptability. While maintaining efficient inference speed, it significantly improves the accuracy of industrial defect detection, basically meeting the strict requirements of real-time quality inspection in aluminum profile production lines. However, in extremely complex scenarios where defects are highly similar to normal textures or multiple defects overlap, the model still has some missed detections. Future research will focus on the deep integration innovation of Transformer and convolutional networks, explore the application of self-supervised learning and domain adaptation technology in expanding industrial datasets, and further enhance the model's generalization ability for complex scenarios, providing better solutions for the development of industrial visual inspection technology.

References

[1] LU J Z, ZHANG Y C, LIU S P, et al. Lightweight DCN-YOLO for Strip Surface Defect Detection in Complex Environments [J]. Computer Engineering and Applications,2023, 59(15):318-328.

[2] WU L, CHU Y K, YANG H G, et al. Sim-YOLOv8 Object Detection Model for DR Image Defects in Aluminum Alloy Welds[J]. Chinese Journal of Lasers,2024,51(16): 29-38.

[3] Li B L. Design and research of automatic cloth sewing machine[D]. Shanghai: Donghua University, 2022.

[4] WU Z H, ZHONG M E, TAN J W, et al. Research on five types of typical defects image detection algorithms for complex textured fabrics[J]. Electronic Measurement Technology, 2023, 46(16): 57-63.

[5] XU Y D, CAI Y H, LI Y, et al. Lightweight overhead transmission line bird's nest detection network based on

YOLOv5s[J]. Electronic Measurement Technology, 2024, 47(7): 138-148.
[6] DONG C, ZHANG K, XIE Z 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): 427-439.
[7] ZENG Q, LI B. Cucumber detection algorithm based on improved SSD[J]. Foreign Electronic Measurement Technology, 2023, 42(04): 158-165.
[8] ZHENG X , SHAO Z ,CHEN Y , et al. MSPB-YOLO: High-Precision Detection Algorithm of Multi-Site Pepper Blight Disease Based on Improved YOLOv8[J].Agronomy,2025,15(4):839-858.
[9] DONG C ,SHEN Y ,FENG Z , et al. Connecting finger defects in flexible touch screen inspected with machine vision based on YOLOv8n[J].Measurement,2025, 17(19)： 246-254.
[10] TU X K, ZHENG S W, YU S H, el at. 3D object detection network based on symmetric shape generation[J]. Chinese Journal of Scientific Instrument, 2023, 44(6): 252-263.
[11] XU F X, FAN R, MA X L. Improved YOLOv7 algorithm for crowded pedestrian detection[J]. Computer Engineering, 2024, 50(3): 250-258.