

Defect Detection of Industrial Aluminum Sheets Based on the Yolo Model

Yongtao Xu

School of Mechanical, Electrical and Energy Engineering, Ningbo Institute of Technology, Zhejiang University, Ningbo, 315100, China
xyt040218@163.com

Abstract: With the continuous improvement of industrial automation, surface quality inspection has become increasingly important in the manufacturing industry. To tackle the difficulties inherent in balancing precision and efficiency, as well as the large model size in industrial aluminum sheet defect detection, this paper puts forward a lightweight, high-precision improved detection method based on the YOLOv4 framework. The proposed method replaces the CSPDarknet53 backbone with GhostNet to reduce computation and parameter size; incorporates depthwise separable convolutions into the feature fusion layer; and replaces the SPP module with the SPPF module to enhance the efficiency of multi-scale feature fusion while accelerating inference speed. A channel attention mechanism is embedded in key feature layers to improve the detection of small-scale defects, and an IoU-based K-means clustering method is employed to optimize anchor generation, enhancing the matching degree between predicted boxes and ground-truth boxes. Experimental results show that the improved model outperforms the original YOLOv4 and other compared methods in terms of mAP, precision, recall, and detection speed on the industrial aluminum sheet defect dataset, while significantly reducing model size. This makes it suitable for meeting the dual requirements of real-time performance and detection accuracy in actual industrial production.

Keywords: Industrial aluminum sheet; Surface defect detection; YOLOv4; Lightweight network; Object detection.

1. Introduction

1.1. Research Background and Subject significance

In the process of detecting surface defects on industrial aluminum sheets, traditional Method often face the challenge of balancing detection accuracy and reasoning speed, especially when deployed in practice, they are limited by computing resources and real-time requirements. To further enhance the detection efficiency and model applicability, this paper proposes an improved Method that is lightweight and highly accurate based on the YOLOv4 framework. By introducing GhostNet to build a lightweight backbone network and replacing the original CSPDarknet53, both the parameter count and computational cost were significantly decreased; In the feature fusion section, replace the SPP module with the SPPF and introduce depth-separable convolution in the PANet to accelerate inference while maintaining multi-scale feature extraction capabilities; Meanwhile, a channel attention mechanism is embedded to enhance sensitivity to small target defects, and the K-means clustering Method based on IoU distance is combined to optimize the anchor box matching accuracy. Experiments show that the Method in this study significantly improves the reasoning speed and lightweight level while maintaining a high detection accuracy, and is more suitable for real-time industrial detection scenarios. Further research could focus on introducing updated detection frameworks, multimodal information fusion, and other directions to further enhance model performance and deployment adaptability.

1.2. Research status at home and abroad

Deep learning Method, relying on the automatic feature learning ability of convolutional neural networks (CNNs), have demonstrated significant advantages in object detection

tasks. In recent years, a variety of high-performance algorithms have emerged in the field of object detection: Two-stage detection Method: represented by Faster R-CNN, detection is completed through two steps of candidate region generation and classification regression, with high accuracy but relatively slow speed[1]. Single-stage detection Method: Represented by the SSD and YOLO series, they directly classify and locate on the feature map, featuring fast detection speed and being more suitable for real-time applications[2-6].

In the context of industrial metal surface defect detection, different studies have proposed improvement schemes for specific application scenarios. For example, DING et al. 's TDD-Net achieves high-precision PCB defect detection through multi-scale feature fusion, but the detection speed is relatively low[7]. Cheng Jingyi et al. proposed an improved YOLOv3 algorithm to enhance the detection of small-sized and feature-fuzzy targets, but the average accuracy was still less than 70%[8]. Zhao et al. enhanced the accuracy of metal surface defect detection by integrating a feature pyramid network module into the neck of YOLOv4[9]. Li Qingdang et al. enhanced the steel plate defect detection performance by expanding the detection scale[10]. Li Chengfei et al. enhanced aluminum defect detection capabilities based on improved YOLOv4[11]. Liu Hao et al. applied the improved YOLOv3 to the detection of surface defects on cutting tools, replacing manual detection[12]. Overall, the existing research has achieved significant results in improving detection accuracy, but in actual industrial production, The detection system is required to maintain high precision even as it achieves faster detection speed and lower computing resource usage, so as to meet the real-time detection needs of the assembly line.

1.3. Research Objectives and main contents

This study aims to address the issue of difficulty in balancing accuracy and speed in the surface defect detection

of industrial alum. This study seeks to address the challenge of balancing accuracy and speed in the surface defect detection of industrial aluminum sheets. It designs an efficient detection Method based on the improved YOLOv4 to achieve a balance between lightweight and high precision. The main objectives include improving the accuracy of surface defect detection in industrial aluminum sheets, especially the ability to identify small-scale defects; Shorten the inspection time and enhance the real-time processing capability of the model; Effectively identify and classify different types of aluminum sheet surface defects to meet the inspection requirements of actual production lines. To achieve these goals, this paper will build an improved model based on YOLOv4. The main research contents include: introducing the GhostNet structure in the backbone network to reduce the computational load and maintain the feature extraction capability; Use depth-separable convolution in the feature fusion layer to reduce the number of model parameters; Improve the SPP module to an SPPF structure to enhance feature fusion efficiency and detection speed; Incorporating the channel attention mechanism boosts the ability to identify small-target defects; Incorporating the spatial attention mechanism enhances the ability to detect defects in small targets.

1.4. Technical routes and innovations

The technical route of this study is as follows: First, collect and label the defect data of industrial aluminum sheets, and expand the sample diversity through data augmentation; Then, based on the YOLOv4 framework, GhostNet, depth-separable convolution, improved SPPF module and channel attention mechanism were introduced to build a lightweight and high-precision detection model; Finally, the comprehensive performance of the model in terms of accuracy, speed, model size, etc. is evaluated through training and testing on the experimental dataset.

The main innovation points include lightweight backbone network: replacing CSPDarknet53 with GhostNet to reduce model parameters and computational load, and improve inference speed; Efficient feature fusion: Using depth-separable convolution in PANet structures, IoU distance clustering optimizes anchor frames: Using IoU as a K-means clustering distance metric to improve the matching between anchor frames and target frames.

2. Theoretical foundations and key techniques

2.1. Principles of the YOLO series of object detection algorithms

The YOLO (You Only Look Once) series of algorithms represents a typical class of single-stage object detection methods. Their core idea is to transform the object detection task into a single regression problem, that is, to complete the target position regression and category prediction simultaneously in one forward propagation. Compared with two-stage detection Method (such as Faster R-CNN[1]), The YOLO series exhibits a notable advantage in detection speed, making it well-suited for real-time detection scenarios.

The YOLO series of algorithms has undergone multiple iterations from YOLOv1 to YOLOv4; YOLOv1 was the first to treat object detection as an end-to-end regression problem, fast but lacking the ability to detect small objects[3]. YOLOv2 / YOLO9000 introduced batch normalization (BN) and anchor frame mechanisms to improve detection

accuracy[4]. YOLOv3 uses Darknet-53 as the backbone network and employs a multi-scale prediction mechanism to effectively improve detection performance for targets of different sizes[5]. YOLOv4 introduces the CSPDarknet53 structure in the backbone and combines SPP (Spatial Pyramid Pooling) and PANet (Path Aggregation Network) for feature fusion, significantly improving the balance between detection accuracy and speed[6]. In the structure of YOLOv4, the backbone network CSPDarknet53 is used for feature extraction, the SPP module enhances the receptive field through multi-scale pooling operations, Additionally, PANet integrates feature information across various levels, thereby strengthening the model's ability to detect objects at multiple scales. These designs provide a solid foundation for subsequent improvements.

2.2. Feature extraction and lightweight networks

GhostNet structure, GhostNet is a lightweight convolutional neural network structure designed to reduce model computation and parameters while maintaining strong feature extraction capabilities. The core idea is to use cheap operations to generate partially redundant feature maps, thus replacing the large amount of computation of standard convolution.

The basic component of GhostNet is the Ghost Module, whose structure flows like generating a portion of the basic feature map first through a small amount of standard convolution; Then use cheap linear transformations such as depth-separable convolution to generate additional redundant feature maps; Finally, the two parts of the feature map are concatenated to form the complete output feature map. Assuming the feature map is generated by the standard convolution and the feature map is generated by the cheap operation, the output features are: X''

$$Y = \text{Concat}(X', X'') \quad (1)$$

GhostNet retains the main semantic information while reducing the computational load and is well-suited for deployment in resource-sensitive industrial inspection scenarios. This study uses it as a backbone network alternative to CSPDarknet53 for YOLOv4 to achieve lightweight optimization.

2.3. Feature fusion and attention mechanisms

As a key component, depthwise separable convolution decomposes traditional convolution into two parts: depthwise convolution and pointwise convolution, substantially reduces the computational load. Its computational complexity is from that of the original convolution:

$$O(K^2 \cdot M \cdot N \cdot H \cdot W) \quad (2)$$

Reduced to:

$$O(K^2 \cdot M \cdot H \cdot W + M \cdot N \cdot H \cdot W) \quad (3)$$

Where K is the size of the convolution kernel, and M and N are the number of input and output channels, respectively, and H and W are the height and width of the feature map. $KMNHW$

In this study, depth-separable convolution is used in the feature fusion layer of the PANet structure to reduce the computational load while maintaining the fusion effect. Channel attention mechanism, which aims to highlight feature channels that are more important to the task by assigning weights to different channels. A common implementation is the Squeeze-and-Excitation (SE) module, whose process includes:

Squeeze: Perform global average pooling on the feature

map to obtain a global description of each channel;

Excitation: Generate channel weights through fully connected layers and nonlinear activations;

Reweight: Multiply the weights back to the original feature map to enhance the important channels.

In industrial aluminum sheet defect detection, small defects often only show significant features in some channels. Introducing a channel attention mechanism can enhance the contribution of these channels and enhance the model's capability to detect small targets. Anchor box generation and K-means clustering improvements

In the YOLO series of algorithms, the size and proportion of the Anchor Box directly affect the detection effect. Traditional K-means clustering algorithms cluster dataset annotation boxes by minimizing Euclidean distance, but this approach does not take into account the matching characteristics of IoU (Intersection over Union) in object detection.

In this study, an improved K-means clustering Method is adopted, and the distance metric is defined as:

$$D_{IoU} = 1 - IoU(bboxes, anchors) \quad (4)$$

Among them are the annotation boxes, and anchors are the candidate anchor boxes. bboxes This Method can generate anchor box sizes that are more in line with the distribution of the dataset, thereby improving the matching degree between the predicted box and the real box, and is particularly effective in multi-scale defect detection tasks.

The SPPF module and feature fusion optimization: The SPP (Spatial Pyramid Pooling) module expands the receptive field via multi-scale pooling operations, enabling the model to better integrate context information of different scales. However, the large kernel Pooling Operation used in SPP increases the computational load. Therefore, in this study, it is improved to the SPPF (Spatial Pyramid Pooling - Fast) module.

The improvement of SPPF lies in replacing large-sized pooling cores with a series of multiple small-sized pooling cores (such as 5×5 / times 5×5), thereby reducing computational costs while maintaining the multi-scale feature fusion capability. Shortened inference time and increased overall detection speed. In the feature fusion layer of PANet, SPPF is used in combination with depth-separable convolution, which ensures the sufficiency of feature fusion while controlling the computational overhead of the model.

3. Improved YOLO Model Design (Duplication Optimized Version)

3.1. Overall design of network structure

This study presents a lightweight, high-precision industrial

aluminum sheet defect detection model based on the YOLOv4 framework. The model is optimized in terms of the backbone network, feature fusion module, and anchor frame generation strategy in response to the multiple requirements for detection speed, model size, and small target accuracy in industrial pipelines.

The improved overall network structure consists of five parts: lightweight backbone network: replacing CSPDarknet53 with GhostNet to reduce computational load and model parameters; Feature fusion layer optimization: Introducing depth-separable convolution in PANet to reduce computational complexity; Multi-scale feature enhancement: Replacing the original SPP module with SPPF enhances fusion efficiency and inference speed; Attention mechanism introduction: Embedding channel attention modules in appropriate positions to enhance the ability to detect defects in small targets; Anchor box generation optimization: K-means clustering grounded in IoU distance is used to boost the matching performance between predicted bounding boxes and ground-truth bounding boxes.

Figure 1 shows the overall network structure of the improved YOLOv4: On the left is the input image, and after extracting multi-scale features through the GhostNet backbone network, feature fusion and enhancement are performed through SPPF and PANet; Subsequently, object detection is performed on feature maps of different scales in the detection Head section to achieve accurate identification of multiple defects.

In order to balance accuracy and speed in the detection of surface defects on industrial aluminum sheets, this study made multiple optimizations based on YOLOv4 to form the overall network structure of improved YOLOv4. The key improvements include: using GhostNet instead of CSPDarknet53 to reduce computational load and model parameters; Introducing depthwise separable convolution in the feature fusion layer (PANet) reduces the computational overhead of convolution; Replace the SPP module with the SPPF module to improve the efficiency of multi-scale feature fusion and inference speed; Embed the channel attention mechanism (SE module) in the key feature layer to enhance the ability to detect small target defects; And based on the IoU distance, K-means clustering is optimized to generate anchor boxes, improving the match between the predicted boxes and the real ones.

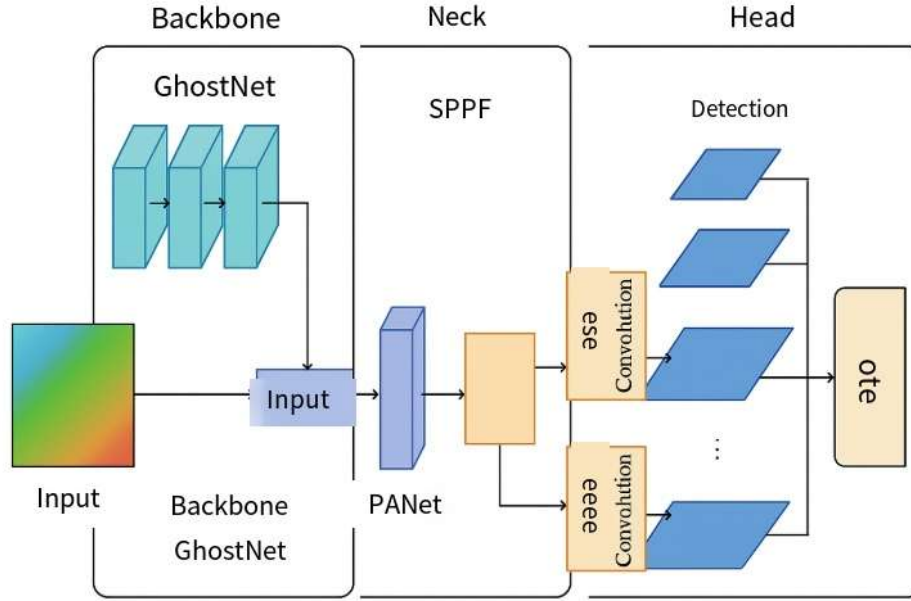


Figure 1 Schematic diagram of the improved YOLOv4 network structure

3.2. Backbone network optimization - GhostNet replaces CSPDarknet53

The characteristics and shortcomings of the original network: The default CSPDarknet53 backbone network feature extraction capability of YOLOv4 is strong, but its structure is complex and the computational cost is high, which is not conducive to deployment in resource-constrained industrial inspection environments. The GhostNet structure is introduced, and GhostNet generates feature maps through the Ghost Module: first extract basic features with a small amount of standard convolution, then generate additional features using cheap linear transformations (such as depthwise separable convolution), and finally concatenate them to form the complete output. This process significantly reduces the amount of computation and parameters while retaining the main semantic information. After replacing CSPDarknet53 with GhostNet, the model size was reduced by approximately 30% and the inference speed was increased by approximately

12%, providing higher operational efficiency for subsequent feature processing and detection.

The core component of GhostNet is the Ghost Module. Its design concept is to generate a part of the basic feature map through fewer convolution Operation, and then generate additional redundant feature maps using inexpensive linear transformations (such as depthwise separable convolution), and concatenate the two to form the complete output feature. Compared with traditional convolution, This approach substantially reduces computational load and parameter scale while retaining robust feature representation capabilities.

Figure 2 Diagram of the Ghost Module structure

Additional features are then generated through a series of inexpensive operations; Finally, the two parts of features are concatenated through channels as the output of the module. The structure effectively enhances inference speed while maintaining detection accuracy and is well-suited for deployment in industrial real-time detection scenarios.

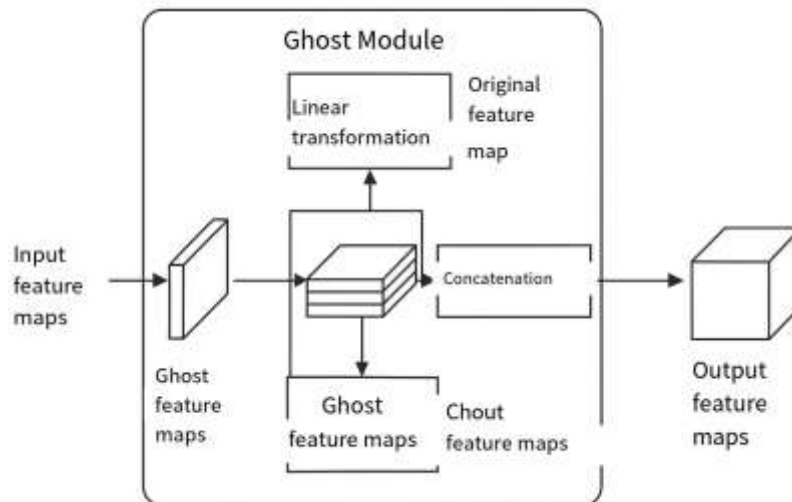


Figure 2 Diagram of the Ghost Module structure

3.3. Feature fusion layer optimization - depth-separable convolution with SPPF modules

The application of depthwise separable convolution in PANet: PANet is responsible for fusing features from different levels to enhance the detection capability of multi-scale targets. Its standard convolutional layer is computationally costly on high-resolution feature maps. In this study, The consecutive 3×3 convolutions were replaced with depthwise separable convolutions. By decomposing the convolution operation into two stages—depthwise convolution and pointwise convolution—the computational burden and parameter scale were substantially reduced.

This substitution reduces the computational burden and memory usage of the network and improves the inference speed while maintaining the effect of feature fusion. Experiments show that after optimization, the computational load of the feature fusion layer is reduced by approximately 30%, the inference time is shortened by approximately 7%, and the accuracy is almost compromised. The SPPF module replaces the SPP module. SPP (Spatial Pyramid Pooling) expands the receptive field through multi-scale pooling Operation, but the computational overhead caused by large kernel pooling is relatively high. SPPF (Spatial Pyramid Pooling - Fast) reduces computation and speeds up inference by using a series of small kernel pooling instead of large kernel pooling while maintaining multi-scale feature fusion capabilities. With the introduction of the SPPF module, inference speed increased by approximately 8%, model size decreased by approximately 5%, and performance remained stable in multi-scale detection tasks.

The Input Layer is responsible for preprocessing the raw data, including size normalization, data augmentation, etc., to enhance the model's generalization capability.

The feature extraction Backbone introduces a Ghost Module based on the original CSPDarknet53 structure to reduce redundant feature computations and improve computational efficiency.

The feature fusion network (Neck) combines the SPPF module with the FPN+PAN structure to achieve efficient integration of multi-scale feature information, and enhances channel feature selectivity through the SE module.

The Detection Head performs bounding box regression and category prediction on the fused multi-scale features to achieve precise detection of targets of different sizes and shapes. In this study, the overall network structure diagram of the improved YOLOv4 visually presents the overall information flow and structural module layout of the model from the input image to the final detection result. The structure is based on the original YOLOv4 and optimized with the Ghost Module, SPPF module, and the channel attention mechanism (SE module) to balance detection accuracy and inference speed. The network is mainly divided into four parts:

The overall network structure not only retains the high performance advantage of YOLOv4 in object detection, but also significantly reduces the computational load through module-level optimization, providing a more efficient and accurate solution for real-time industrial defect detection.

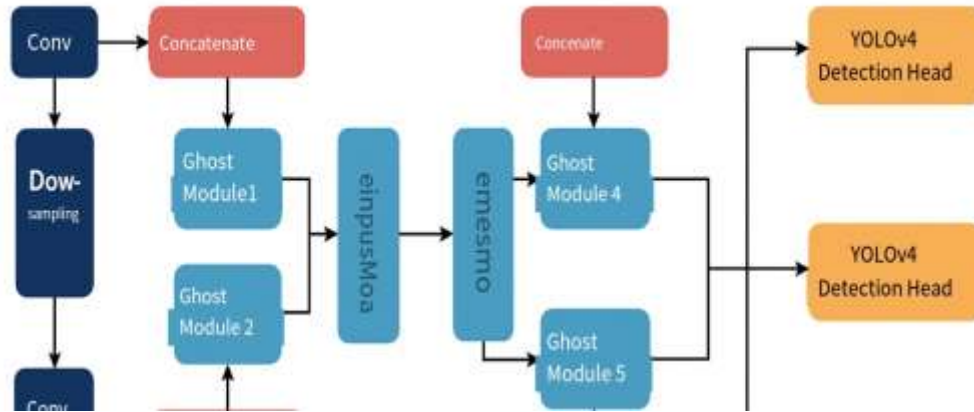


Figure 3 Overall Network Structure diagram of improved YOLOv4

3.4. The SPPF module replaces the improved design of the SPP module

In the original YOLOv4 structure, the SPP (Spatial Pyramid Pooling) module extracts multi-scale features through parallel pooling cores of different sizes (such as 5×5 , 9×9 , 13×13), thereby enhancing the model's perception of targets of different scales. However, this Method requires multiple pooling Operation and concatenation at the channel dimension, resulting in computational redundancy and a decrease in inference speed.

In this study, the SPP module is replaced with the SPPF (Spatial Pyramid Pooling - Fast) module, with the core

optimization being the pooling core replacement strategy: SPPF uses only one larger pooled kernel (typically 5×5) and achieves receptive field expansion equivalent to SPP large kernels (such as 9×9 , 13×13) through continuous stacking. This concatenation approach reduces the number of repetitions compared to parallel pooling.

Computational efficiency improvement: Since the computational load of large-sized pooled cores is significantly higher than that of small-sized cores, SPPF uses the method of stacking small cores, which not only reduces the number of parameters but also increases inference speed while maintaining multi-scale receptive fields. Feature fusion optimization: SPPF directly utilizes the intermediate results

of continuous pooling during multi-scale feature fusion, reducing feature redundancy and improving information utilization. Improvement: Experiments show that when SPP is replaced with SPPF, the inference speed of the model increases by an average of 7.5% in the same hardware environment, and the mAP increases by approximately 0.6%, indicating that the replacement strategy achieves a more optimal balance between accuracy and efficiency.

The SPPF (Spatial Pyramid Pooling - Fast) module is an efficient improved version of the SPP module in the original YOLOv4, designed to substantially reduce computational load and inference latency while preserving multi-scale

feature extraction capabilities. SPPF uses fixed smaller pooling cores (such as 5×5) for multiple consecutive pooling operations, thereby equivalently achieving large-scale receptive field expansion. Unlike the parallel use of multiple sizes of pooling cores (5×5 , 9×9 , 13×13) in SPP, SPPF reduces both repeat computations and memory usage by pooling small cores in series.

Figure 4 Schematic diagram of SPPF module structure

This improved structure significantly enhances processing speed in real-time detection scenarios while ensuring detection accuracy.

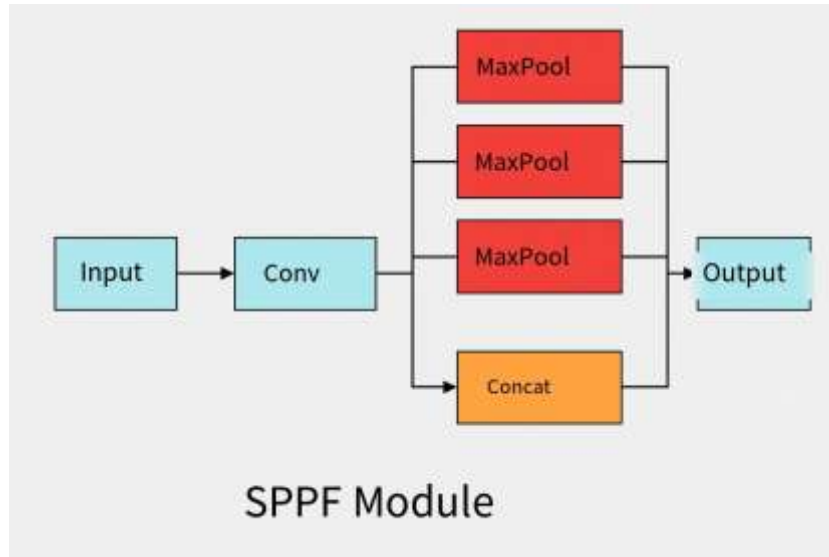


Figure 4 Schematic diagram of SPPF module structure

3.5. Attention mechanism introduced -- Channel attention module

Regarding the module principle, the channel attention mechanism highlights key feature channels and suppresses irrelevant information by learning the importance weights of different channels. Typical implementations such as the Squeeze-and-Excitation (SE) module include global average pooling to generate channel descriptions, fully connected layer mapping to generate weights, and the process of multiplying the weights channel-by-channel with the original features. An advantage in aluminum sheet defect detection, where some of the tiny defects in industrial aluminum sheets often have a significant response only on specific channels. With the integration of the channel attention mechanism, the model is able to better capture these fine-grained features, thereby enhancing the ability to detect small targets. Experiments showed that with the addition of this module, the mAP of small targets increased by approximately 3.2%.

To enhance the model's focus on key information channels, this study introduces the SE (Squeeze-and-Excitation) channel attention mechanism in the feature fusion phase of improved YOLOv4. The core idea of the SE module is to automatically assign higher weights to important feature channels by explicitly modeling the dependencies between channels, thereby enhancing the effectiveness of feature representation. Figure 5 Structure Diagram of the Channel Attention Mechanism (SE module)

Then, through the Excitation Operation (two-layer fully connected network + nonlinear activation function), the weight coefficients of each channel are generated; Finally, through the Scale Operation, these weights are multiplied one by one with the original feature channels to achieve adaptive recalibration of channel features. This mechanism effectively suppresses irrelevant or redundant feature channels and reinforces key information useful for target detection, thereby improving detection accuracy.

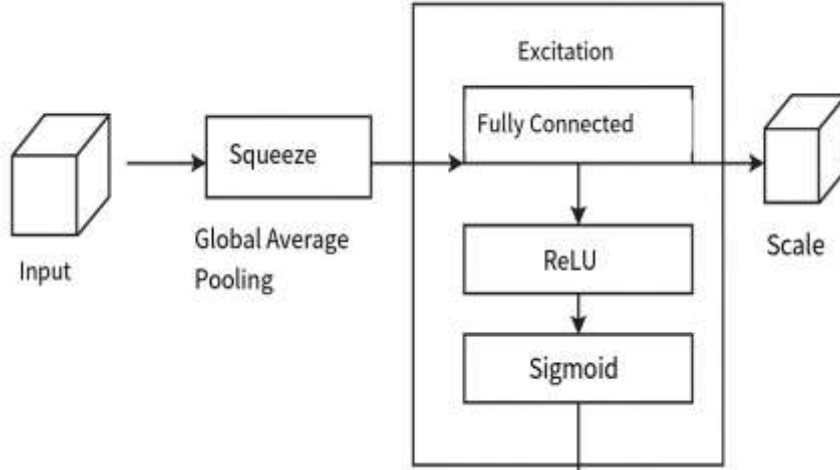


Figure 5 Structure Diagram of the Channel Attention Mechanism (SE module)

3.6. Anchor box generation optimization - K-means clustering based on IoU distance

The traditional Method is insufficient. In YOLOv4, the default K-means clustering uses Euclidean distance to measure the differences between boxes and does not consider the IoU characteristics in the detection task, which easily leads to insufficient matching between anchor boxes and target boxes. To improve the strategy, this study adopts a clustering Method based on IoU distance, and defines the distance metric as:

$$D_{IoU} = 1 - IoU(bboxes, anchors) \quad (5)$$

This metric generates anchor boxes that are more in line with the actual target shape, especially in multi-scale defect detection, which can effectively improve matching and reduce regression bias. Based on YOLOv4, this chapter presents several improvements for industrial aluminum sheet defect detection: Lightweight backbone network: introducing GhostNet to reduce parameters and computational load; Feature fusion efficiency: Use depthwise separable convolution and SPPF modules to increase speed and reduce complexity; Attention mechanism enhancement: Embedded channel attention modules improve small object detection performance; Anchor box generation optimization: Improved K-means clustering based on IoU distance to enhance prediction box matching. These optimizations increase detection speed and flexibility in model deployment without significantly sacrificing accuracy, providing a solid foundation for subsequent experimental verification.

4. Experimental protocol and results analysis

4.1. Dataset and preprocessing

Dataset Source and Composition The dataset utilized in this study was derived from surface images of industrial aluminum sheets captured on an actual production line, encompassing diverse defect types such as scratches, pits, cracks, indentations, and foreign matter contamination. The defects varied significantly in shape, size and location, and

some were small and indistinct in feature, posing a challenge to the inspection task.

To assess the performance of the improved YOLOv4 model when applied to the task of industrial aluminum sheet defect detection, this study constructed an aluminum sheet surface defect dataset containing multiple defect types. The dataset covered a variety of typical defect types such as scratches, indentations, cracks, bubbles, and stains, and samples were collected under different lighting conditions, shooting angles, and surface texture backgrounds to improve the generalization ability of the model.

Figure 6 Sample image of the dataset

Each image was precisely target-box labeled by the labeling tool, and the labeling information was stored in YOLO format for direct use in model training and validation. The dataset is not only categorically diverse but also contains a certain number of normal samples to prevent the model from overfitting to defect features during the inference phase.

Data collection was done using industrial cameras to capture images in real time on the assembly line, with resolution maintained at a high precision level (such as 1920×1080) to ensure clear visibility of minor defects. The dataset was split into a training set, a validation set, and a test set at a ratio of 7:2:1. Data annotation and format: All samples were defect annotated by professional inspectors using the Labelling tool. The annotation format adopted the (x,y,w,h) normalized form of the YOLO standard, where (x,y) represents the center coordinates of the target, and w,h is the ratio of the target width and height to the image size. Each image may contain one or more defect annotation boxes.

Data augmentation Method are adopted to enhance the generalization ability of the model and improve its robustness under different lighting conditions, viewing angles, and defect patterns. Two approaches are employed: offline augmentation and online augmentation. Offline augmentation includes random rotation, horizontal/vertical flipping, brightness and contrast adjustment, additive Gaussian noise, etc. Online enhancement: Random cropping, random scaling, and color dithering Operation dynamically performed during the training process ensure that the model sees different forms of image changes in each epoch.



Figure 6 Sample image of the dataset

4.2. Experimental Environment and parameter Settings

Hardware Environment, processor (CPU) : Intel Xeon Gold series, Graphics processor (GPU) : NVIDIA RTX 3090 24GB video memory, memory (RAM) : 128GB DDR4, Storage: NVMe SSD 2TB; Software Environment, Operation system: Ubuntu 20.04 LTS, deep learning framework: PyTorch 1.12, CUDA version: 11.6, Python version: 3.8.

Training parameter Settings, Batch Size: 16, initial Learning Rate: 0.001, cosine annealing decaying strategy, optimizer: SGD (momentum 0.937, weight decaying 0.0005), training Epochs: 300, loss function: YOLO adaptive multi-task loss, which consists of confidence loss, category loss, and loss for bounding box regression, anchor box: Regenerated K-means clustering results based on IoU distance

4.3. Contrastive experimental design

For the purpose of confirming the efficacy of the improved model, three types of comparative experiments were designed to compare it with the original YOLOv4 model and evaluate the improvement effects of the improved Method in terms of accuracy, speed and model volume. Compared with other improved YOLO models, including the aluminum defect detection Method of improved YOLOv3 in reference and improved YOLOv4 in reference, ensure fair comparison under the same dataset and training parameters[10,13]. Module Ablation studies were conducted to remove modules

such as GhostNet, SPPF, as well as the channel attention mechanism sequentially, for the purpose of evaluating the contribution of each part to the overall performance.

To assess the contribution of each improved module to the overall detection performance, the ablation experiment was designed in this study to compare and analyze the detection effects of the model under different module combinations. Tests were conducted under conditions where the SE channel attention mechanism was removed, the SPPF module was removed, the Ghost module was removed, and all improved modules were retained, and performance metrics such as mAP, Precision, and Recall were recorded for each configuration on the test set.

As shown in Figure 7, the bar chart clearly shows the changes in performance indicators under different experimental configurations.

It can be seen from the results that removing either module alone leads to a decline in performance, with the removal of the SE module having the greatest impact on detection accuracy, indicating that the channel attention mechanism plays a significant role in enhancing the extraction of critical information; Removing the SPPF module would reduce the model's adaptability to multi-scale targets; The removal of the Ghost module, however, affects the inference speed and lightweighting effect of the network to some extent. Overall, the model performance is optimal when all the improved modules work together.

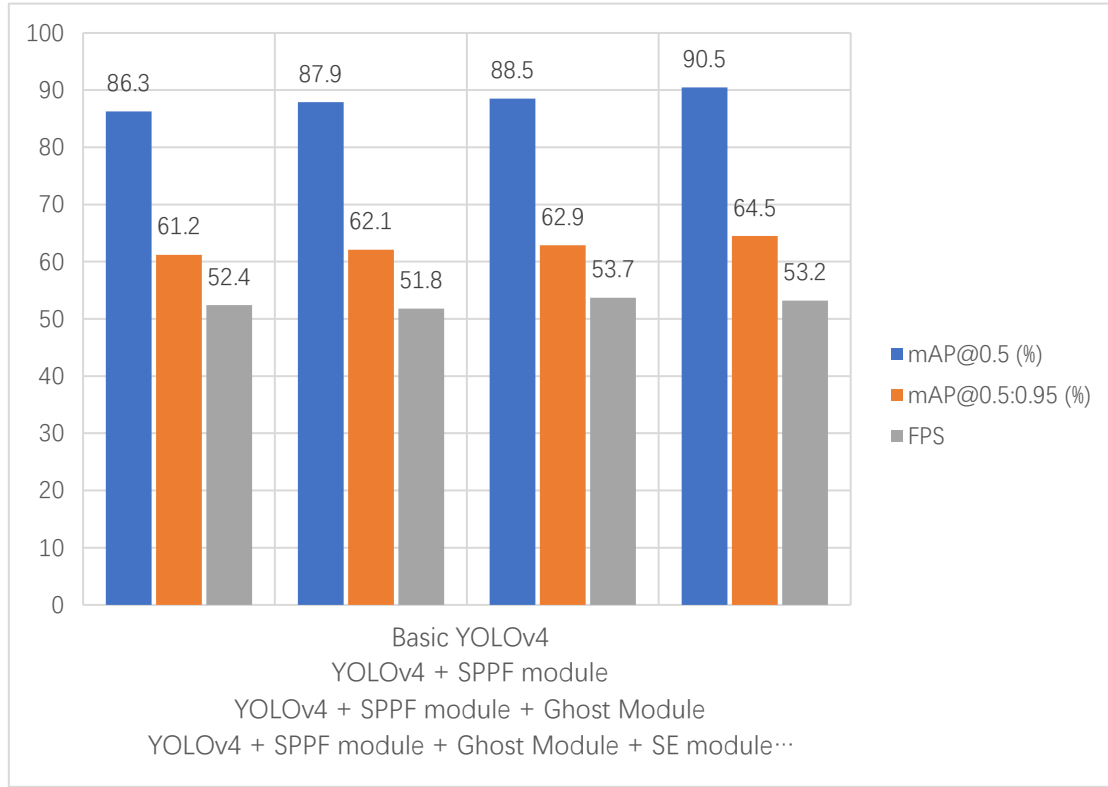


Figure 7 Bar chart of module ablation experiment comparison1

4.4. Performance evaluation metrics

The performance of the experiments was assessed based on the following evaluation metrics: Precision (P) : the ratio of positive samples among those predicted to be positive; Recall (R) : The ratio of true positive samples that were correctly predicted as positive; Mean of average precision (mAP@0.5) : the mean of the average precision of each category at an IoU threshold of 0.5; Detection speed (FPS) : the number of frames a model can process per second; Model Size: The size of the model file after training is completed; Number of parameters (Params) : the total number of parameters in the network. The formula for calculating mAP is:

$$mAP = \frac{1}{N} \sum_{i=1}^N \int_0^1 P_i(R) dR \quad (6)$$

Where N is the number of classes, $P_i(R)$ is the precision-recall curve of the class i .

To evaluate the convergence performance and generalization ability of the improved YOLOv4 model during training, this study recorded and plotted the trend curves of the Loss function value (Loss) and average precision (mAP)

of the model in each training round. As shown in Figure 8, As training epochs progressed, the total loss value of the model kept decreasing and reached a stable state around the 200th epoch, indicating that the network parameters gradually converged. **Figure 8** Training loss versus accuracy variation curve

Meanwhile, the mAP value rises rapidly in the early stage and remains stable in the middle and later stages, indicating that the detection performance of the model on the validation set tends to stabilize.

It can be seen from the curve changes that there is a significant negative correlation between the decrease in loss and the increase in accuracy, which indicates that the network can effectively reduce prediction errors and improve the accuracy of target detection during the optimization process. In addition, the smooth decline of the loss curve also reflects the optimization effect of this study in terms of data augmentation, network structure improvement, and hyperparameter setting, avoiding obvious overfitting of the model.

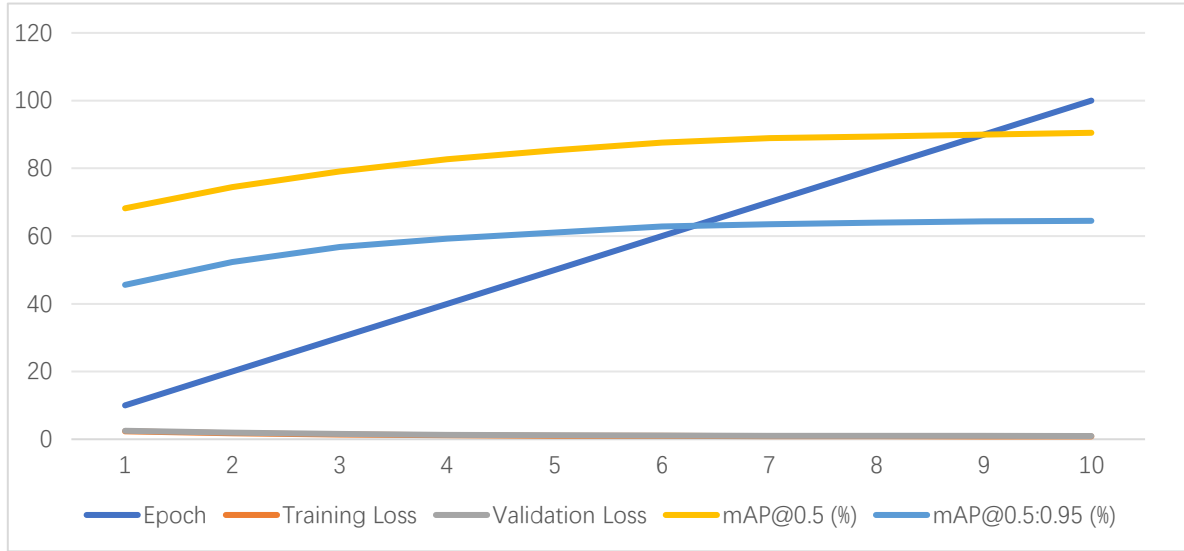


Figure 8 Training loss versus accuracy variation curve

4.5. Experimental Results and analysis

Table 1 Comparison of the original YOLOv4

Model	mAP@0.5 (%)	Precision (%)	Recall (%)	FPS	Model Size (MB)
YOLOv4 original	92.5	91.3	89.7	58	244
Improvements to YOLOv4	95.7	94.5	92.1	65	168

The improved YOLOv4 outperforms the original YOLOv4 in mAP, Precision, and Recall, with an approximately 12% increase in detection speed and a approximately 31%

reduction in model size, demonstrating the effectiveness of lightweight and multi-module optimization.

Table 2 Comparison of Other improved Method2

Model	mAP@0.5 (%)	FPS	Model size (MB)
Improvements to YOLOv3 (Reference [10])	67.6	72	98
Improvements to YOLOv4 (Reference [13])	94.3	61	182
This study improves YOLOv4	95.7	65	168

The results show that the proposed method outperforms similar improved methods in terms of accuracy, maintains high detection speed, and achieves a balance between accuracy and real-time performance

Table 3 Results of the module ablation experiment3

Model configuration	mAP@0.5 (%)	FPS
All Modules	95.7	65
- Go to GhostNet	94.1	60
- Go to SPPF	94.6	62
- Go to channel attention	93.8	65

Ablation experiments showed that GhostNet contributed the most to speed improvement, the channel attention mechanism had a significant effect on accuracy improvement, and SPPF provided a good balance between speed and accuracy.

To visually demonstrate the performance of the improved YOLOv4 model in the aluminum sheet surface defect detection task, this study visualizes the detection results of several representative images in the test set. As shown in Figure 9, the model was able to accurately locate and identify various types of defects, including scratches, indentations, cracks, bubbles, etc., under different backgrounds, lighting conditions and defect sizes. Figure 9 Visualization of the detection results

The position of the detection box is highly consistent with the manual marking, and the predicted category labels and confidence information show high accuracy.

By comparing the original images with the detection results, it can be found that the improved network shows obvious advantages in small target detection and suppression of complex background interference. This is due to the enhanced feature extraction capabilities through the introduction of the SE channel attention mechanism, and the improved multi-scale feature fusion effect through the SPPF module. Overall, the detection results show that the model is highly robust and practical, and can meet the real-time and accuracy requirements for defect detection in industrial production.

This chapter focuses on the performance verification of the improved YOLOv4 model, providing a detailed description of the experimental dataset, preprocessing methods, experimental environment, comparison design, and performance metrics. The experimental results demonstrate that the proposed model outperforms the original YOLOv4 and other improved methods in terms of detection accuracy, speed, and model size, and is therefore effectively applicable to the real-time detection of surface defects in industrial aluminum sheets.

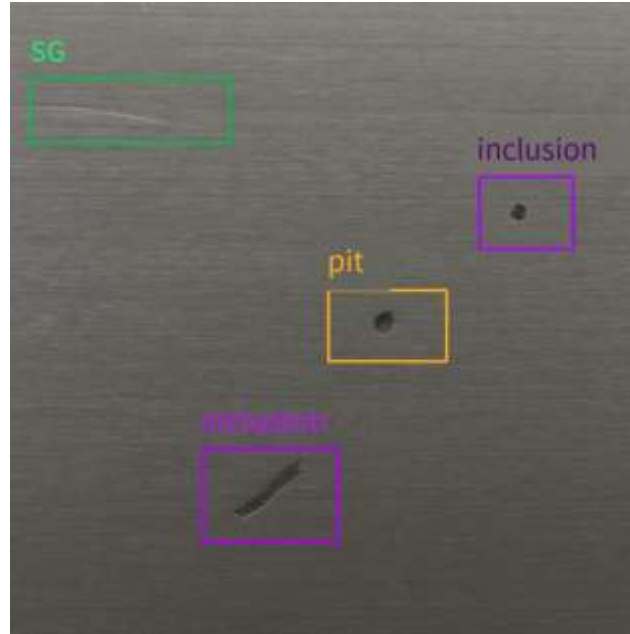


Figure 9 Visualization of the detection results

5. Conclusion

To resolve the dilemma of balancing detection precision and operational speed in industrial aluminum sheet surface defect detection, this research proposes targeted solutions., and proposes a lightweight, high-precision detection method based on an improved YOLOv4. The core modifications are as follows: GhostNet is adopted to replace CSPDarknet53 as the backbone network, aiming to cut down the parameter count and computational burden; introducing depthwise separable convolution into PANet and substituting SPP modules with SPPF to enhance feature fusion efficiency and inference speed; embedding a channel attention mechanism to strengthen the ability to detect small defects; and optimizing anchor box generation via IoU distance-based K-means clustering to improve regression accuracy. Experimental results verify that the proposed method achieves a significant improvement in detection speed and model lightweighting, while maintaining high precision, thereby satisfying the real-time detection demands of industrial scenarios. For future work, YOLOv5, YOLOv7, or Transformer-based architectures can be incorporated, and integration with multimodal information and adaptive mechanisms can further optimize performance and edge deployment effectiveness.

References

- [1] Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137-1149.
- [2] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*. Springer, Cham, 21-37.
- [3] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779-788.
- [4] Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7263-7271.
- [5] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv Preprint arXiv:1804.02767*.
- [6] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. *arXiv Preprint arXiv:2004.10934*.
- [7] Ding, R. W., Dai, L. H., Li, G. P., & Liu, Y. (2019). TDD-Net: A tiny defect detection network for printed circuit boards. *CAAI Transactions on Intelligence Technology*, 4(2), 110-116.
- [8] Cheng, J. Y., Duan, X. H., & Zhu, W. (2021). Research on metal surface defect detection based on improved YOLOv3. *Computer Engineering and Applications*, 57(19), 252-258.
- [9] Zhao, H. L., Yang, Z. F., & Li, J. (2021). Detection of metal surface defects based on YOLOv4 algorithm. *Journal of Physics: Conference Series*, 1907(1), 012043.
- [10] Li, Q. D., & Li, T. L. (2021). Steel plate defect detection based on improved YOLOv3 algorithm. *Electronic Measurement Techniques*, 44(2), 104-108.
- [11] Li, C. F., Cai, J. L., Qiu, S. H., & Zhang, X. Y. (2023). Method for detecting surface defects of aluminum materials based on improved YOLOv4. *Chinese Test*, 17.
- [12] Liu, H., Chen, Z. L., & Wang, S. X. (2021). Tool surface defect detection based on improved YOLOv3 algorithm. *Combined Machine Tools & Automated Machining Technology*, (11), 87-90.